Evaluación de la calidad de las aguas superficiales mediante técnicas de estadística multivariante: Un estudio de caso en la cuenca del Río Paute, al sur de Ecuador

Tesis presentada en cumplimiento parcial de los requisitos para la obtención del grado de Magister en Ecohidrología, Facultad de Ingeniería y Facultad de Ciencias Naturales y Museo, Universidad Nacional de La Plata, Argentina

Por:

Gonzalo Sotomayor

La Plata – Argentina

2016

Nombre del Director:	Dra. Henrietta Hampel
Nombre del Codirector:	Ing. MSc. Pablo Romanazzi
Fecha de defensa de tesis:	5 de agosto de 2016

AGRADECIMIENTOS

Al Dr. Piercosimo Tripaldi de la Universidad del Azuay, por su estupenda supervisión y ayuda durante el desarrollo de este proyecto, especialmente en la aplicación de las diferentes técnicas de estadística multivariante.

Estoy muy agradecido con mi directora y codirector de tesis, Dra. Henrietta Hampel de la Universidad de Cuenca, e Ing. MSc. Pablo Romanazzi de la Universidad Nacional de La Plata, por su apoyo y orientación brindados durante el desarrollo de este estudio.

Un agradecimiento muy especial al Ing. Juan Pablo Martínes Romero e Ing. Marco Jaramillo Valdivieso, ex miembros de la SECRETARÍA DEL AGUA – DEMARCACIÓN HIDROGRÁFICA SANTIAGO; ambos compañeros durante su gestión apoyaron incondicionalmente la idea y desarrollo de este estudio.

A los Doctores Jorge Emilio Celi Sangurima de la Michigan State University y Raúl Acosta Rivas de la Universidad de Barcelona, gracias por su meritoria ayuda en el planteamiento y diseño iniciales de este trabajo, sus consejos y comentarios me ayudaron muchísimo.

A la Universidad Nacional de La Plata, a su Facultad de Ciencias Naturales y Museo y a su Facultad de Ingeniería, ambas gestoras del Programa de Maestría en Ecohidrología, gracias por su formación académica constante.

Al Milano Chemometrics and QSAR Research Group del Dipartimento di Scienze dell'Ambiente e del Territorio e di Scienze della Terra, Università degli Studi di Milano – Bicocca y al Natural History Museum de la University of Oslo, por la apertura en el uso de sus herramientas informáticas 'classification toolbox' y 'ga toolbox' para MATLAB y PAST: PALEONTOLOGICAL STATISTICS SOFTWARE PACKAGE 2.17, respectivamente.

A mis entrañables amigos argentinos, Esther, Raquel, Hugo y Juan, gracias siempre por su soporte imperecedero, por su bondad y afecto; estas personas durante mi estancia en Argentina me hicieron sentir como en casa.

A Rominita, gracias por estar. Y asimismo, por su valioso aporte en la edición del presente documento.

A mi familia, especialmente a mí abnegada madre, padre y hermanas, gracias por su amor y su apoyo incuestionables.

Gonzalo Sotomayor
Agosto 2016

In God we trust, all others bring data.

- William Edwards Deming (1900 - 1993)

TABLA DE CONTENIDOS

Lista de Figuras

Lista de Mapas

Lista de Tablas

Lista de Abreviaturas

ABI	Andean Biotic Index
BMWP	Biological Monitoring Working Party
CCA	Canonical correspondence analysis
CG Paute	Comité de Gestión de la Cuenca del Paute (Ecuador)
CREA	Centro de Reconversión Económica del Azuay, Cañar y Morona Santiago
CRP	Cuenca del río Paute (Ecuador)
DHS	Demarcación Hidrográfica Santiago (Ecuador)
EPT	Efemeróptera, Plecóptera, Trichoptera
EH	Ecohidrología
FSS	Final stepwise selection (Etapa de selección final)
GAs	Genetic Algorithms (Algoritmos Genéticos)
IRHA	Instituto de Recursos Hídricos del Azuay (Ecuador)
k – NN	*k* - Nearest Neighbour
MDT	Modelo digital del terreno
MR	Multiple regressions
OCH	Odonata, Coleóptera, Heteróptera
PCA	Principal Components Analysis
SC	Sistema Complejo
SENAGUA	Secretaría del Agua
UDA	Universidad del Azuay (Ecuador)

RESUMEN

Diferentes técnicas de estadística multivariante como el método de clasificación del vecino más cercano (*k* – NN) a través de algoritmos genéticos (GAs), un análisis de componentes principales (PCA) y regresiones múltiples (MR), se llevaron a cabo para evaluar e interpretar bajo el marco conceptual de la Ecohidrología una gran y compleja matriz de datos de calidad de agua. Los datos se obtuvieron durante cinco años (2008, 2010 - 2013) de muestreo en la cuenca del Río Paute al sur de Ecuador. Treinta y cuatro variables físico - químicas, microbiológicas, geomorfológicas y biológicas (macroinvertebrados bentónicos) fueron monitoreadas en 64 sitios (10234 observaciones). El análisis *k* - NN a través de GAs se utilizó para conocer cuál de 6 índices bióticos dados por los macrozoobentos es, en términos de ajuste matemático para un modelo de clasificación, el óptimo. Se obtuvo como resultado que una combinación de puntajes del Andean Biotic Index (ABI; zonas > 2000 m.s.n.m.) y el Biological Monitoring Working Party calibrado para Colombia (BMWP/Col; zonas < 2000 m.s.n.m.) es la variable de respuesta biológica más adecuada. Una redistribución de las clases de los índices bióticos mostró que matemáticamente estas se optimizan si son tres (dadas por los percentiles 33,33 % y 66,66 % del índice biótico) y no cinco. Se aplicó un PCA sobre tres grupos de calidad de agua establecidos por los percentiles 33,33 % y 66,66 % del índice biótico seleccionado (combinación ABI + BMWP/Col), siendo las variables que mayoritariamente explican a las mejores clases y su estado de buena integridad ecológica (clase 1) la presencia de vegetación de bosque de ribera y la alta heterogeneidad del lecho. Por el contrario, elevados niveles de coliformes fecales, demanda bioquímica de oxígeno (DBO), amonio, turbiedad, pH y temperatura del agua, más bajas valoraciones de calidad de hábitat; son condiciones que se asocian con clases de aguas contaminadas (clase 3). Finalmente, un método de validación para los resultados del PCA basado en Regresiones Múltiples se probó con éxito enfatizando así la fiabilidad científica del estudio.

Palabras Clave: Estadística multivariante, Ecuador, cuenca del río Paute, respuesta biológica.

ABSTRACT

Multivariate statistics techniques such as *k* - Nearest Neighbor Method (*k* - NN) with genetic algorithms (GAs), principal component analysis (PCA) and multiple regressions (MR), were applied in order to evaluate and interpret, from an Ecohydrological viewpoint, a large complex matrix of water quality data collected during five years (2008, 2010-2013) at the Paute river basin (southern part of Ecuador). Thirty four physical, chemical, microbiological, geomorphological and biological (benthic macroinvertebrates) variables at 64 different sites (10234 observations) were assessed. A first approach was carried out to identify which of the six biotic indices obtained through benthic macroinvertebrates is the optimum for a mathematical adjustment in a classification model (*k* - NN by GAs). As a result, the combination of the Andean Biotic Index scores (ABI; elevation > 2000 meters above the sea level – m.a.s.l) and the Biological Monitoring Working Party calibrated for Colombia (BMWP / Col, elevations < 2000 m.a.s.l) was regarded as being the most appropriate biotic index (biological response variable). A redistribution of the classes of biotic indexes showed that, mathematically, these are optimized when the classification system has three rather than five classes, expressed by percentiles (33,33 % and 66,66 %). PCA was applied at three groups of water quality data defined by percentiles of the most appropriate biotic index (ABI + BMWP / Col). The variables that explain the status of good ecological integrity (class 1) are the presence of riparian vegetation and the high heterogeneity of streambed. Contaminated water was related to levels of fecal coliform, biochemical oxygen demand, ammonium, turbidity, pH and temperature, and was classified (bioassessed) as bad habitat (class 3). Finally, an independent method based on multiple regressions for validation of results of the PCA was successfully tested, emphasising the scientific reliability of the study.

Key Words: Multivariate statistics, Ecuador, Paute river basin, biological response variable.

1. INTRODUCCIÓN

La calidad de las aguas superficiales es un tema muy delicado. Influencias antrópicas tales como el desarrollo de las urbes, industrias, actividades agrícolas, el aumento del consumo de los recursos hídricos, degradan a las mismas y perjudican su funcionalidad y optimización en los usos domésticos, industriales, agrícolas, creativos u otros. Así como ciertos procesos naturales, por ejemplo los cambios en los regímenes de precipitación, la erosión y el desgaste de los materiales de la corteza terrestre (Carpenter et al., 1998). En contexto, estas tendencias se consideran ya amenazas a una escala global (Strobl & Robillard, 2008), por lo tanto es imperativa la necesidad de una eficiente y adecuada gestión a largo plazo de los cuerpos de agua superficiales para lo que se requiere una comprensión fundamental de las características hidromorfológicas, químicas y biológicas de los mismos.

Sin embargo, todo esto se complica debido a que las variaciones espaciales y temporales en la calidad del agua son muy difíciles de interpretar. Por ello, un programa de monitoreo que proporcione una estimación representativa y fiable de la calidad de las aguas superficiales, es fundamental en el manejo y gestión de las mismas (Dixon & Chiswell, 1996).

Empero, en algunas ocasiones los programas de monitoreo y vigilancia para la calidad del agua incluyen altas frecuencias de muestreos en muchos sitios y asimismo la determinación de un gran número de parámetros físico – químicos, microbiológicos, etc. Bajo estos antecedentes, normalmente como resultado se obtiene una matriz de datos de gran tamaño, que de por sí se constituye en un sistema complejo (SC) y que necesita, bajo este precepto, una interpretación con técnicas de análisis de datos para un SC (Chapman, 1992). Con ello, en general es sabido que los sistemas ambientales son muy confusos con una red de interacciones químicas, físicas y biológicas que determinan sus características. Por lo tanto, para caracterizar las propiedades de una cuenca u otro ecosistema definible, se necesitan grandes conjuntos de datos multivariantes. Y así se puede proporcionar las bases e identificar las fuentes de determinados compuestos y la comprensión de su transporte, transformación y destino. Estos grandes conjuntos de datos se prestan idóneos para análisis matemáticos y estadísticos que los transformen en resultados más compactos e interpretables, y que pueden servir como base para descripciones ordenadas de un sistema así como para el desarrollo de los modelos matemáticos deterministas que son muy necesarios para estimar los efectos potenciales de las futuras actividades o estrategias de control.

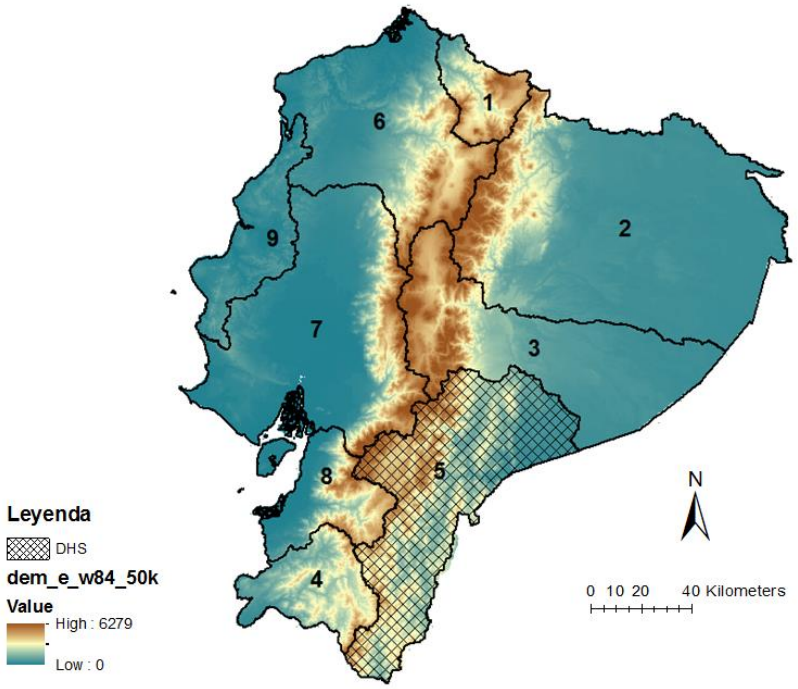
A saber, la aplicación de diferentes técnicas de estadística multivariante, como ciertos análisis de ordenamiento o métodos de clasificación, ayudan en la interpretación de las complejas matrices de datos para entender mejor la calidad del agua y el estado ecológico de los sistemas estudiados. Estos métodos permiten la identificación de los factores / fuentes que influyen mayoritariamente en los sistemas hídricos y ofrecen una valiosa herramienta para la gestión fiable de recursos. Como consecuencia de ello, muchas técnicas de estadística multivariante se han aplicado para caracterizar y evaluar la calidad del agua dulce superficial (Vega et al., 1998; Wunderlin et al., 2001; Simeonov et al., 2003; Ouyang, 2005; Shrestha & Kazama, 2007).

Todos estos antecedentes no son en lo absoluto ajenos a la realidad de Ecuador. De tal modo, en el presente estudio una gran matriz de datos obtenida durante un programa de monitoreo de 5 años (2008 – 2010 – 2013) se sometió a diferentes técnicas de estadística multivariante para extraer: (1) la información acerca de las similitudes o diferencias entre los sitios de muestreo; (2) para evidenciar cuál índice dado por macroinvertebrados bentónicos es el óptimo en valorar el estado ecológico de los ríos y quebradas; (3) identificar las variables responsables de las variaciones espaciales en la calidad del agua de los ríos y (4) examinar los

factores ocultos que explican la estructura de la base de datos sobre los parámetros de calidad del agua de la Cuenca del río Paute (CRP) ubicada al sur de Ecuador.

2. PROBLEMÁTICA Y JUSTIFICATIVO

Los programas de monitoreo de calidad del agua surgen a nivel mundial como respuestas de tipo iniciales, previas a planes de manejo o programas de restauración ambiental y su objetivo es el de proporcionar una estimación representativa y fiable de la calidad de las aguas considerando las variaciones espacio – temporales. Todo esto con el único fin de ser lo más precisos (variables adecuadas a evaluarse) y exactos (réplicas representativas a nivel espacio - temporal) en cuanto a confiabilidad de muestreo se refiere. De tal modo, todas estas circunstancias están presentes y se dan en la actualidad en muchas zonas de Ecuador, así, en la CRP se han venido desarrollando desde hace algunos años, varios trabajos cuyo propósito ha sido el de levantar información de calidad de agua a nivel superficial. En contexto y por un orden e importancia institucional, es preciso señalar que estas actividades fueron llevadas a cabo en su momento por entidades que en la actualidad ya no existen, tales como el Consejo de Aguas de la Cuenca del Paute (CGPaute), el Instituto de Recursos Hídricos del Azuay (IRHA) y el Centro de Reconversión Económica del Azuay, Cañar y Morona Santiago (CREA). Por circunstancias intrínsecas al Ecuador que ha experimentado cambios abruptos en las diversas políticas de desarrollo regional y local desde el año 2008, se creó la Secretaria del Agua (SENAGUA), la cual absorbió a todas las instituciones involucradas en el manejo y gestión de recursos hídricos y se constituyó desde entonces como la autoridad única del agua en todo el territorio nacional. La SENAGUA dentro de sus primeras acciones de gestión dividió al país en 9 demarcaciones a nivel de Ecuador continental (Mapa 1), las cuales no obedecen a un orden político sino más bien a uno hidrológico. La zona del presente estudio, la CRP, pertenece a la Demarcación Hidrográfica Santiago (DHS) (Mapa 1).



Mapa 1. Demarcaciones hidrográficas del Ecuador continental (nueve). Se detalla la Demarcación Hidrográfica Santiago (DHS) (5) a la cual pertenece la CRP.

Desde el año 2008 en la CRP se han realizado varios trabajos efectuados sobre todo por universidades locales, cuyo fin ha sido el de establecer estados de calidad del agua que sirvan para efectivizar el manejo y gestión de las aguas superficiales. Variables de tipo físico –

químicas, microbiológicas y biológicas (macroinvertebrados bentónicos identificados a nivel de familia) han sido evaluadas en diferentes épocas de los distintos años en muchas estaciones ubicadas a lo largo y ancho de la CRP. Esto parece lo ideal, empero, en los siguientes párrafos se detallan en un orden jerárquico los pormenores de la problemática existente y que fue abordada en el presente estudio.

Al ser trabajos efectuados por entidades ya inexistentes, mucha de la información de calidad de agua se encontraba dispersa y aún no se había logrado reunir en un solo compendio por parte de la SENAGUA – DHS, esto significaba una grave NO OPERATIVIDAD y NO FUNCIONALIDAD de la información levantada. Asimismo, muchas actividades como los monitoreos futuros se deben de planificar y redefinirse (en caso de ser necesario) en base a estudios predecesores lo cual no se llevó a cabo. De tal modo, la tarea de unificar la información se puso en marcha desde marzo de 2014.

Sin embargo, desde un principio se observó que el gran conjunto de datos de calidad del agua no estaba completo, en otras palabras, no había una secuenciación adecuada entre un estudio y otro, de tal modo que:

a. Existían estaciones cuya representatividad en el tiempo fue de solamente una réplica, y por el contrario, otros puntos de monitoreo se caracterizaban por poseer varias repeticiones de muestreo.
b. Ciertas variables evaluadas estaban presentes en un estudio, mientras que en otros no.

Los puntos **a** y **b** fueron claves, ya que impulsaron la organización de la información en una sola matriz con datos empatados o equiparados a nivel espacial - secuencial - temporal. Dicha matriz se compone de 10234 campos constituidos por 64 estaciones de muestreo más sus réplicas (301 objetos en total) y 34 variables de tipo físico – químicas, microbiológicas, geomorfológicas y biológicas (macrozoobentos). Dada la naturaleza de la matriz lograda, el componente temporal no es objeto de análisis en este estudio sino únicamente el espacial.

3. OBJETIVO DEL ESTUDIO

Desagregar parte de la información contenida en la gran y compleja matriz de datos de calidad de agua y así, sentar las bases para en un futuro crear adecuadas herramientas de gestión que permitan optimizar el manejo y gestión de los recursos hídricos superficiales en la CRP.

4. PREGUNTAS A RESPONDER E HIPÓTESIS

Enmarcados en el enfoque Ecohidrológico (EH) y dada la naturaleza de la problemática que se ha descrito, se plantearon las siguientes preguntas:

- P_1: ¿Existen diferencias, entre distintos índices bióticos dados por la comunidad de macrozoobentos, en términos de cuál de ellos es la respuesta biológica para calidad de agua óptima respecto de varios descriptores físico – químicos, microbiológicos y geomorfológicos?
- P_2: ¿Qué variables de tipo físico – químicas, microbiológicas y geomorfológicas son más importantes para describir o explicar la variabilidad espacial de la calidad del agua dada por los macroinvertebrados bentónicos en la CRP?
- H_1: Los ríos tropicales son descritos con mucha frecuencia por su biodiversidad, y básicamente por la presencia o no de taxas y la tolerancia de estas a procesos contaminantes. En base al criterio de las distintas resoluciones taxonómicas que se emplean, se supone que hay diferencias entre los múltiples índices dados por los macrozoobentos (unos poseen más y otros menos capacidad de respuesta biológica).

Se cree que estas diferencias podrían evidenciarse en términos de un buen o mal ajuste matemático (dependiendo del índice) en un modelo de clasificación. Por tal motivo se propuso investigar cuál de varios índices estudiados es el adecuado a nivel regional para la CRP.

- H_2: Se cree que existen variables del tipo descriptoras (físico – químicas, microbiológicas y geomorfológicas) que explican con mayor peso o importancia la variabilidad de los macroinvertebrados bentónicos, mientras que otras posiblemente tengan una influencia menor sobre esta variable de respuesta biológica.

5. MARCO CONCEPTUAL DEL ESTUDIO: ECOHIDROLOGÍA (EH)

Está bien establecido que los ecosistemas acuáticos (arroyos, ríos, estuarios, lagos, humedales y entornos marinos) están estructurados por la interacción de procesos físicos, químicos y biológicos en múltiples escalas tanto a nivel espacial como temporal (Frothingham et al., 2002; Thoms & Parsons, 2002; Dauwalter et al., 2007). De tal modo, la necesidad de una investigación interdisciplinaria para hacer frente a las múltiples interrogantes científicas se ha reconocido cada vez más (Dollar et al., 2007) y ha dado lugar a la aparición de nuevas "sub-disciplinas" para hacer frente a estos cuestionamientos (Hannah et al., 2007).

Es así que la Ecohidrología (EH) es una de estas áreas de investigación emergentes que ha unido a biólogos, ecólogos, geomorfólogos, sedimentólogos, hidrólogos, ingenieros hidráulicos y fluviales para hacer frente a las preguntas de investigación fundamentales que harán avanzar los temas de ciencia y gestión integral claves para mantener los ecosistemas acuáticos y atender a las demandas que impone la sociedad contemporánea sobre los mismos (Maddock et al., 2013). La EH se ha desarrollado en la interface permeable de las disciplinas tradicionales, combinando el estudio de las propiedades y procesos hidrológicos asociados con la típica ingeniería hidráulica y la geomorfología y su influencia en la ecología acuática y la biología de organismos (Vogel, 1996). En contexto, Porter & Rafols (2009) sugirieron que los desarrollos interdisciplinarios de la ciencia han sido mayores entre las disciplinas estrechamente aliadas y menos desarrollados para los campos con una mayor distancia entre ellos, en tal virtud, se conjetura a la EH como un campo y concepto eficiente y en constante desarrollo (Fig. 1).

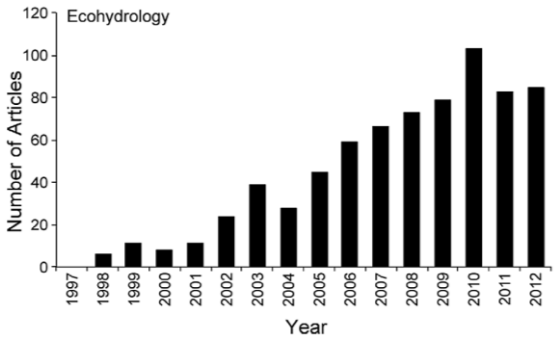


Figura 1. Artículos publicados para ecohidrología, eco-hidrología o eco hidrología en el periodo de 1997-2012 (Maddock et al., 2013)

El concepto de EH se desarrolló en el marco del Programa Hidrológico Internacional de la UNESCO PHI-V (Zalewski et al, 1997) y se ha inspirado en gran medida por las conclusiones de la Conferencia Internacional sobre el Agua y el Medio Ambiente (ICWE) celebrada en Dublín en 1992. Esta conferencia puso de manifiesto la insuficiencia de las soluciones existentes en las prácticas de gestión del agua para lograr la sostenibilidad de los recursos hídricos. Así como la necesidad de nuevos conceptos y soluciones (Zalewski, 2002), desde entonces, muchos aportes han sido un cúmulo de evidencia para la sostenibilidad de la EH en el tiempo. A saber, una adecuada conceptualización de la EH radica en que esta es una ciencia interdisciplinar, que

busca comprender los vínculos entre los factores de estrés físico - químicos con receptores biológicos para así apoyar las políticas para la sostenibilidad de los recursos naturales (Loinaz, 2012). De forma macro, los estudios referidos a EH pretenden arrojar como resultado un "estado ecológico" producto de las modificaciones en el ambiente (Fig. 2).

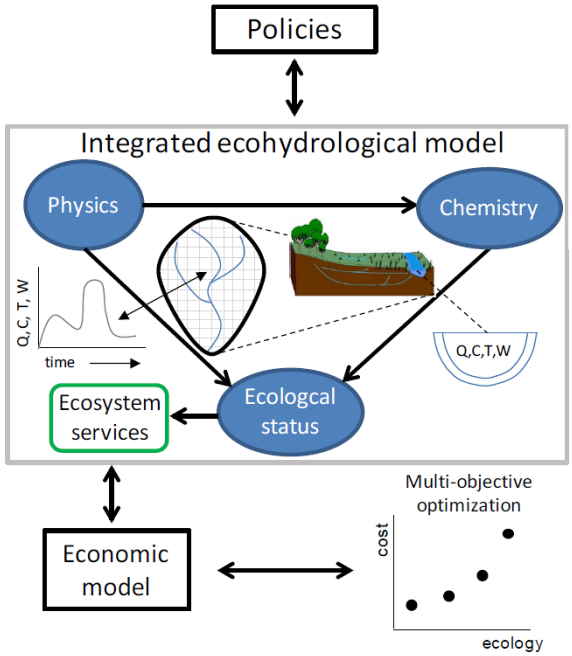


Figura 2. Marco conceptual para la EH integrada en el contexto de la gestión de los recursos hídricos. Q = caudal, C = concentración de solutos, T = temperatura del agua, W = peso de peces u otros organismos (Loinaz, 2012).

Con estos antecedentes, el marco conceptual bajo el cual se realizó el presente trabajo fue la EH, siendo los macroinvertebrados bentónicos nuestra variable de repuesta biológica respecto de varias descriptoras del tipo físico – químicas, microbiológicas y geomorfológicas.

6. ESTADO DEL ARTE (métodos de estadística multivariante empleados para responder las preguntas científicas planteadas)

La estadística multivariante es considerada como el mejor método para evitar malas interpretaciones de las grandes y complejas matrices de datos que resultan como producto de los programas de vigilancia ambiental (Juahir et al., 2010). Estos métodos han sido ampliamente utilizados en la elaboración de información significativa a través de masas de datos ambientales. A menudo se han empleado como herramientas exploratorias de datos para la clasificación de las muestras (observaciones) o estaciones de muestreo y la identificación de las fuentes de contaminación. Es por ello que la estadística multivariante también se ha aplicado para caracterizar y evaluar la calidad del agua dulce, así como la verificación de las variaciones espaciales y temporales causadas por factores naturales y antrópicos.

A continuación se describen los métodos que se utilizaron en este estudio para responder las dos preguntas de investigación planteadas.

6.1. La clasificación multivariante

Durante los últimos 25 años, ha existido un gran esfuerzo para mejorar las metodologías analíticas aplicadas a problemas de investigación a nivel químico, biológico y ambiental. Asimismo, en cualquiera de estas áreas de estudio es necesario analizar un gran volumen de datos (como es el presente caso) para evaluar y acceder a la amplia variación (información) que un compuesto químico, biológico o sistema ambiental puede poseer. No obstante, la información presente en estas grandes y complejas matrices no se puede extraer si los datos se examinan de forma unitaria para cada variable. Por lo que las relaciones investigadas en estos conjuntos de datos por lo general no se pueden expresar en términos cuantitativos sino a nivel de similitud o diferencia entre grupos de datos multivariados. Es por ello que la tarea que corresponde a la investigación de las relaciones en este tipo de datos multidimensionales es doble:

- ¿Puede una estructura de datos multivariados basados en subgrupos de muestras distintas ser identificada?.
- Una muestra desconocida ¿Puede ser clasificada en uno de estos subgrupos para la predicción de una propiedad de la muestra?.

Para los investigadores estas dos cuestiones se han abordado mediante la clasificación o técnicas de reconocimiento de patrones supervisados (Tauler et al., 2009), los cuales son nombres dados a un conjunto de técnicas numéricas desarrolladas para resolver el problema de asignar un objeto a una clase; siendo la formulación del problema la siguiente:

“Dada una colección de muestras caracterizadas por un conjunto de mediciones realizadas en cada una de ellas, el objetivo es encontrar y / o predecir una propiedad de las muestras que no sea directamente medible en sí mismo o sea muy difícil de estimar, pero que se conoce está indirectamente relacionada con las mediciones a través de alguna correspondencia desconocida o no determinada” (Sharaf, 1986).

En un estudio de reconocimiento de patrones las muestras se clasifican de acuerdo a una propiedad específica utilizando medidas que indirectamente están relacionadas con esa propiedad. Una regla de clasificación se desarrolla a partir de un conjunto de muestras para las que se conoce la propiedad de interés y sus medidas; luego, la regla de clasificación se utiliza para predecir esta propiedad en muestras que no son parte del conjunto de entrenamiento inicial. El conjunto de muestras para el que la propiedad de interés y las medidas se conocen se llama el conjunto de entrenamiento (training set), mientras que el conjunto de medidas que describen cada muestra en el grupo de datos se denomina patrón. La determinación de la propiedad de interés mediante la asignación de una muestra a su respectiva clase se llama reconocimiento, de ahí el término "reconocimiento de patrones” (Tauler et al., 2009). Las técnicas de reconocimiento de patrones pueden dividirse en dos grandes categorías:

- Las discriminantes que usan una medida de distancia para identificar la clase a la cual un objeto o muestra esta ligado en el espacio de patrones.
- Métodos basados en particiones que clasifican una muestra dividiendo el espacio de datos en diferentes regiones[1].

[1] En el caso más simple, un clasificador binario, el espacio de datos se divide en dos regiones, las muestras que comparten una propiedad común se encuentran en un lado de la superficie de decisión, mientras que aquellas muestras que comprenden la otra categoría se encuentran antagónicas a las primeras (Tauler et al., 2009).

6.2. Funciones lineales discriminantes

En un estudio de reconocimiento de patrones, cada muestra u objeto en el grupo de datos está representado inicialmente como un vector de datos $X = (x_1, x_2, x_3, \dots, x_j, \dots x_P)$ en donde el componente x_j es una variable medible en la muestra, y cada una de ellas se considera como un punto en un espacio de medición p - dimensional. La dimensionalidad del espacio de medición corresponde con el número de variables utilizados para caracterizar cada muestra u objeto. El supuesto básico es que las distancias Euclidianas[2] entre los pares de puntos en este espacio de medición están inversamente relacionadas con el grado de similitud entre las muestras correspondientes, tal es así, que puntos que representan objetos de una clase se agruparán en una región limitada del espacio de medición distante de los puntos correspondientes a otra clase.

A modo de introducción a los clasificadores lineales, se muestra la Fig. 3, en donde es evidente que las dos clases pueden ser convenientemente separadas por una línea, siendo $d(x) = w_1x_1 + w_2x_2 + w_3 = 0$ la ecuación de la recta o la frontera (límite) de decisión, en donde $w_{1\dots k}$ equivale al peso de la combinación lineal de las variables que caracterizan a los objetos y x_1 , x_2 son las coordenadas de variables para cada muestra en el conjunto de datos. Entonces, dado un patrón vector $\boldsymbol{X}$ se puede decir que x pertenece a la Clase 1 si $d(x) > 0$ ó a la clase 2 si $d(x) < 0$. De todo esto también es importante señalar que una combinación lineal de dos variables originales cumple con el propósito de análisis mientras que una medición individual de x_1 y x_2 no serviría para este propósito.

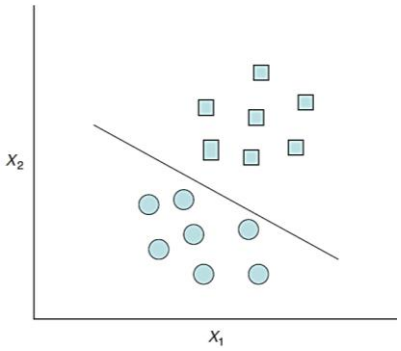


Figura 3. Clasificador binario simple. Los cuadrados representan muestras de la clase 1 y los círculos representan las muestras de la clase 2 (Tauler et al., 2009).

Cualquier muestra puede clasificarse respecto a una superficie discriminante lineal calculando el producto escalar del vector patrón aumentado y el vector de peso. La distancia entre la superficie de decisión y cada muestra está dada por la magnitud del producto de punto, que también se llama la puntuación discriminante. El signo del producto escalar denota el lado de la superficie de decisión en la que se encuentra la muestra. Para las muestras en un lado de la superficie de decisión, los productos de puntos o puntuaciones discriminantes son siempre positivos mientras que los patrones en el lado opuesto tienen puntuaciones discriminantes negativas (Sin embargo si el producto escalar es cero, es falso concluir que la muestra se encuentre en la superficie de decisión). Por lo tanto, cualquier muestra en el conjunto de datos se puede clasificar en una de las dos categorías mediante la obtención de la señal de la puntuación discriminante. Para un problema de clasificación, donde cada muestra se caracteriza por dos mediciones o variables, la superficie lineal de decisión tomará la forma de una línea, mientras que la superficie lineal de decisión será un plano si cada muestra se caracteriza por tres variables.

[2] Distancia Euclidiana (Hammer, 2012). $d_{jk} = \sqrt{\sum_i (x_{ji} - x_{ki})^2}$

La superficie lineal de decisión es un hiperplano si el número de mediciones utilizados para caracterizar cada muestra en el conjunto de datos es mayor que tres.

En contexto, las funciones lineales discriminantes se enmarcan en dos categorías:

1. Métodos paramétricos o probabilísticos
2. Métodos no paramétricos o no probabilísticos.

Los métodos paramétricos (1) se basan en la estadística Bayesiana y dependen de tener las funciones de densidad de probabilidad de las clases o las estimaciones de ellas; usan los vectores de medias y matrices de covarianza de las dos clases de base para el desarrollo y el centrado de la superficie de la clasificación. Por el contrario, si las propiedades estadísticas de las clases no pueden ser calculadas o estimadas, o simplmente no se desea someterse a principios estadísitcos demasiado rígidos que quizás no se cumplen en el mundo real, se utilizan los métodos no paramétricos (2) (Varmuza, 1980). Estos métodos generan análisis discriminantes basados en el conocimiento de valores de pertenencia de clase y de datos sin la necesidad de utilizar la información sobre las medidas estadísticas de sus distribuciones, dicho de otro modo, no necesitan hipótesis previas sobre la distribución de las variables.

Así, ejemplos de técnicas no paramétricos son el *k* – NN (***k*** - Nearest Neighbor), o método del vecino mas cercano, el cual es el algoritmo que en este estudio se utilizó para responder a las preguntas científicas planteadas, principalmente a la primera.

6.3. El método del vecino mas cercano (*k* – NN)

Es un algoritmo de clasificación no paramétrica y uno de los métodos más antiguos y simples de reconocimiento de patrones de aprendizaje automático (Niemann, 1983; Cunningham & Delany, 2007; Tauler et al., 2009), aunque dicha simpleza no lo exime para ser una de las técnicas de estadística multivariante más poderosas. En la Conferencia Internacional de Minería de Datos (ICDM) celebrada en diciembre de 2006 el *k* – NN fue identificado dentro de los 10 algoritmos de minería de datos más influyentes en la comunidad de investigación (Wu et al., 2008).

Este algoritmo fue desarrollado en 1961, en promejora de los análisis discriminantes para cuando las estimaciones confiables sobre las funciones de densidad de probabilidad de las clases eran desconocidas o difíciles de calcular (Kutzer, 2008). Las propiedades formales del *k*-NN se elaboraron por parte de Cover & Hart, (1967); las investigaciones con enfoques sobre la distancia ponderada por Dudani, (1976); Bailey & Jain, (1978); y los métodos difusos por Jóźwik, (1983); Keller et al. (1985).

A saber, hay tres elementos claves en el enfoque de *k* – NN (Wu et al., 2008):

- Un conjunto de objetos etiquetados, por ejemplo, un conjunto de registros almacenados. Estas etiquetas, clases o grupos pueden definirse por (1) conocimiento previo (teoría, o evidencias experimentales), (2) interpretación de los resultados de un cluster análisis o (3) discretización de respuestas cuantitativas (Todeschini, 1998).
- Una distancia o similitud métrica para calcular la distancia entre los objetos.
- El valor de *k*, que es el número de vecinos más cercanos.

Así, en el *k*- NN una muestra u objeto se clasifica de acuerdo al voto de la mayoría de sus *k*-vecinos más cercanos. Para una muestra dada, un tipo de distancia (generalmente se aplica la Euclidiana) se calcula a otro punto en el conjunto de datos y estas distancias se disponen de menor a mayor para definir cada objeto. Sobre la base de una etiqueta de clases de la mayoría de las muestras a estas se les asigna una clase en el conjunto de datos, si la clase asignada y la

real coinciden, la prueba se considera un éxito. La tasa de éxito global de clasificación calculada sobre todo el conjunto de datos, es una medida que muestra el grado de éxito en la correcta asignación de objetos o muestras a una clase (Tauler et al., 2009).

En síntesis, para clasificar un objeto sin etiqueta, la distancia de este objeto a los objetos etiquetados se calcula, sus *k* vecinos más cercanos son identificados, y las etiquetas de clase de estos vecinos más cercanos sirven entonces para determinar la etiqueta de clase del objeto (Wu et al., 2008).

Este enfoque de la clasificación es de particular importancia en la actualidad debido a que los problemas de bajo rendimiento en el tiempo de ejecución no son ya significativos con la potencia de cálculo que está disponible gracias a los nuevos ordenadores cuya capacidad de cálculo crece prácticamente de manera exponencial. En síntesis, este proceso tiene dos pasos, el primero es para determinar el número de vecinos cercanos y el segundo para establer la clase usando a estos vecinos (Cunningham & Delany, 2007).

En contexto, para ilustrar la forma en que el algoritmo del *k* - NN trabaja, la Fig. 4 incluye un esquema en el cual la muestra a ser evaluada (círculo verde) debe clasificarse ya sea en la primera clase (cuadros azules) o en la segunda clase (triángulos rojos). De tal modo, si *k* = 3, el objeto se clasifica a la segunda clase porque hay 2 triángulos y sólo 1 cuadro dentro del círculo interior, por el contrario si *k* = 5, la asignación se da en la primera clase (3 cuadros vs. 2 triángulos dentro del círculo exterior).

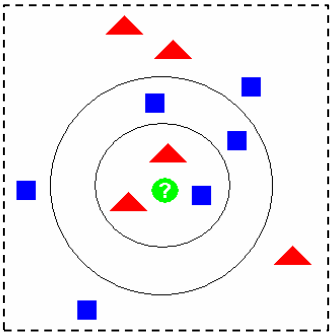


Figura 4. Esquema que detalla la forma en que el algoritmo del *k* - NN trabaja (Kutzer, 2008).

En el ejemplo que se detalla (Fig. 4) resulta una asignación de patrones muy evidente, sin embargo hay más atenciones no tan simples que determinan factores que deben considerarse para la óptima selección de *k*. En el siguiente ejemplo (Fig. 5), se puede ver el efecto del valor "*k*" en los límites de cada clase; se grafican las diferentes fronteras que separan las dos clases (rosa y azul) con diferentes valores de *k*. Se observa que la frontera se vuelve más suave con el aumento del valor de *k* y mientras ésta tienda a aumentar hacia el infinito, los objetos a ser clasificados finalmente serán asignados todos en la clase azul o todos en la roja.

Con estas referencias, el valor óptimo de *k* debe determinarse adecuadamente y una opción para elegirlo es por medio de procedimientos de validación cruzada, ensayando un conjunto de valores de *K* (por ejemplo, del 1 al 10); entonces, el valor de *k* que brinde el menor error de clasificación en la validación cruzada se puede seleccionar como óptimo.

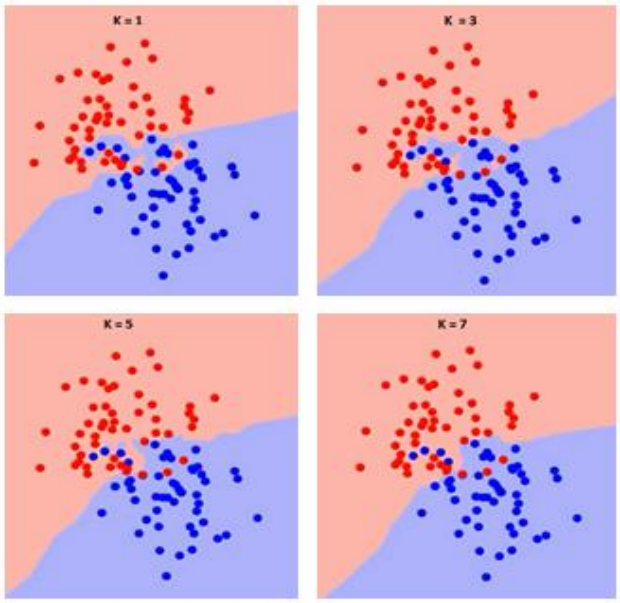


Figura 5. Ilustración del efecto del valor "*k*" (http://www.analyticsvidhya.com).

Con estos antecedentes, se constata que hay varias cuestiones clave que afectan el desempeño del *k* – NN y que se deben de tomar muy en consideración:

- Como se ha enfatizado, la elección del valor de *k* es crucial. Si *k* es demasiado pequeño, entonces el resultado puede ser sensible a los puntos de ruido. Por otra parte, si su valor es demasiado grande, entonces el barrido puede incluir demasiados puntos de otras clases; *k* puede definirse como el parámetro de suavizado (Sahigara et al., 2013).
- La elección del tipo de distancia. Aunque diversas medidas pueden usarse para calcular la distancia entre dos puntos, la más deseable es aquella para la cual una menor distancia entre dos objetos implica una mayor probabilidad de tener la misma clase (para ambos objetos).
- Atributos de variables. Es bien sabido que medidas de distancia como por ejemplo la Euclidiana se vuelven menos discriminantes a medida que el número de atributos aumenta, es por ello que técnicas de escalado o normalización de datos deben tomarse en cuenta la mayoría de veces.
- El *k* - NN es considerado dentro de la categoría de "lazy learners", es decir, "aprendices lentos". Los modelos no se construyen de forma explícita, por lo tanto, la construcción del modelo implica un costo bajo. Sin embargo, la clasificación de objetos desconocidos es relativamente complicada, ya que se requiere el cálculo de los *k* - vecinos más cercanos del objeto a etiquetar (aunque con los nuevos ordenadores la potencia y tiempo de cálculo son mucho más óptimas).

Finalmente, la calidad de la clasificación se evalúa a través de la matriz de confusión; en esta matriz las líneas representan las clases verdaderas, y en la diagonal principal se reportan los objetos clasificados correctamente por el algoritmo y fuera de esta los objetos mal clasificados. Asimismo el parámetro más simple y más usado que sintetiza la calidad del procedimiento es la tasa de error (Error rate % o ER %) o su función inversa, el porcentaje de clasificaciones correctas (Non - Error Rate = NER %) (Todeschini, 1998; Hand, 2012). El ER no es más que el porcentaje de objetos mal clasificados. De la misma forma, la especificidad, sensibilidad y precisión de cada clase ayudan a evaluar el contexto global de la clasificación.

En síntesis, el algoritmo del *k* - NN es:

a. Escalado de los datos

a.1: Generalmente se utiliza Autoscaling[3] de manera que todas las variables están representadas en la misma medida

b. Selección del tipo de medida de distancia

b.1: La más común es la distancia Euclidiana

c. Selección del número de vecinos (*k*)

c.1: Se varía el valor del número de vecinos hasta alcanzar el mejor poder predictivo

d. Cálculo de la matriz de distancias

e. Para cada objeto se evalúa la frecuencia de cada clase entre los vecinos

f. Se asigna el objeto a la clase con más frecuencia

6.4. *k* - NN en combinación con Algoritmos Genéticos (GAs)

A pesar de que para muchos investigadores el *k* – NN logra un rendimiento muy bueno en sus experimentos con diferentes conjuntos de datos, ciertas limitaciones asociadas a la aplicación de este método han sido identificadas:

1. La dependencia en el conjunto de entrenamiento: el clasificador se genera sólo con las muestras de entrenamiento y no utiliza ningún otro dato adicional. Esto hace que el algoritmo dependa marcadamente del conjunto de entrenamiento, y se necesita un nuevo cálculo, incluso si hay un pequeño cambio en el conjunto de entrenamiento.
2. No hay diferencia de peso entre las muestras: no hay diferencia entre las muestras con un pequeño y un gran número de datos. Por lo tanto, y sobre todo para sistemas ambientales, esto no es apropiado dada la aleatoriedad de los sistemas complejos (SC) donde las muestras comúnmente tienen una distribución desigual.
3. Los algoritmos de clasificación, en general, pueden bajar en su rendimiento con un mayor número de características que describan los objetos a ser clasificados.

Con el propósito de solucionar los problemas que causarían estas limitantes, una combinación del *k* + NN más algoritmos genéticos (GAs) fue propuesta (Suguna & Thanushkodi, 2010) (*k* – NN + GAs). En este sentido en el algoritmo tradicional *k* - NN, inicialmente la distancia entre todas las muestras de prueba 'training set' y de capacitación 'test' se evalúan y los *k* - vecinos con mayores distancias se consideran para la clasificación, sin embargo en el *k* – NN + GAs un *k* - número de muestras se eligen para cada iteración y la precisión más alta se registra en cada ocasión, por lo tanto, no se requiere calcular las similitudes entre todas las muestras, y tampoco hay necesidad de considerar el peso de la categoría (Suguna & Thanushkodi, 2010). Lo que pretende el *k* - NN + GAs es dar una solución de optimización a un problema en clasificación, básicamente, (1) efectuando las posibles combinaciones para dar soluciones a un problema de clasificación; y (2) memorizando las mejores soluciones posibles y así, paulatinamente llegar a un óptimo modelo de clasificación.

[3] Con el Autoscaling todas las variables quedan representadas en la misma medida, con media = 0 y Desviación estándar = 1 (Todeschini, 1998). $X_{ij} = \frac{X_{ij} - \overline{X_j}}{S_j}$

En contexto, para ilustrar mejor este proceso dado por el *k* – NN + GAs se debe ahondar en los principios de los GAs, a saber (Ison et al., 2005):

Los GAs corresponden a la clase de métodos estocásticos de búsqueda. Mientras la mayoría de estos métodos operan sobre una única solución, los GAs operan en una población de soluciones. La idea básica, inspirada en los procesos evolutivos en biología, es que el contenido genético de una población contiene potencialmente la solución, o una solución mejor, a un dado problema de adaptación. Esta solución puede estar inactiva porque la combinación genética adecuada está diseminada entre varios sujetos. Sólo la asociación de genomas distintos puede llevar a la activación de la solución.

Crudamente, el mecanismo evolutivo procede así: sobre una población, algunos individuos son seleccionados para la reproducción, con más oportunidades para los mejor adaptados al ambiente. Durante la reproducción, los nuevos individuos de la población resultan de modificaciones e intercambio genético de los padres. Una vez que se renueva la población, el proceso recomienza. Es decir que hay dos espacios donde opera la evolución. Por una parte, a nivel de los individuos físicos (fenotipo), que deben adaptarse para ser seleccionados. Y luego, a nivel de la información genética (genotipo), a través de los operadores que intercambian y varían la información genética.

La información genética está codificada en los cromosomas, que son secuencias de genes, cada uno de los cuales codifica una característica particular del individuo. Estas secuencias son escritas en términos de cuatro bases nitrogenadas: adenosina, timina, citosina y guanina. En este alfabeto de base cuatro, [A; T; C; G], radica escrita toda la información genética de un individuo.

Hay esencialmente dos operadores genéticos. El operador de mutación introduce cierta aleatoriedad en la búsqueda simplemente cambiando unos genes por otros, contribuyendo a una exploración azarosa en el espacio genético. El operador de crossover, en cambio, es una recombinación de la información durante la reproducción de los individuos seleccionados.

El proceso de evolución, puesto en estos términos, es adaptable a una enorme familia de problemas, incluso ajenos al ámbito biológico.

En términos crudos, la meta de la exploración genética es encontrar los individuos mejor adaptados a su ambiente. Para eso, los individuos se reproducen buscando, con el intercambio de material genético y las mutaciones, que cada nueva generación mejore la adaptación. Para poder aplicar este esquema a un problema de 'asignación de patrones ' se debe dar las definiciones de individuos, genes, cromosomas y ambiente, y cuantificar la adaptación.

Si pensamos a cada individuo de una población biológica como un objeto (o estación de muestreo para nuestro caso) como un ente en el espacio multidimensional, por ejemplo (x; y) ϵ [a; b] x [c; d], (en nuestro caso es mucho más complejo que esto pues tenemos $x, y, \dots 33_k\ variables$) podemos definir la adaptación de éste a la minimización de la función $f(x\ , y)$.

Bajo el marco conceptual de un modelo de clasificación, la adaptación se define como la correcta asignación de estaciones de muestreo (cada una descrita por muchas variables físico – químicas, microbiológicas y geomorfológicas) en clases que para nuestro caso específico fueron determinadas por los índices bióticos evaluados.

Para establecer la codificación genética de la estación de muestreo y aplicar los operadores genéticos, definimos un cromosoma como el arreglo consecutivo de dos genes, para el caso de (x; y) ϵ [a; b] x [c; d] es uno para cada número del grupo de variables x, y. Este arreglo se

construye normalizando cada número según el rango donde puede variar y guardando los primeros $\boldsymbol{n}$ decimales. Por ejemplo, para el par (0,5; 1,34) ϵ [0; 1] x [0; 2,35], la normalización (en base al máximo posible) arroja el par (0,5 / 1; 1,34 / 2,35) = (0,5; 0,57021276…). La identificación del individuo con su cromosoma resulta usando cuatro cifras significativas (0,5; 1,34) → [50005702].

En este `espacio genético' se pueden aplicar los operadores de cruzamiento y mutación, que en la evolución suceden en el espacio de las bases nitrogenadas y, aquí, en la base decimal. Una mutación será el reemplazo de cualquiera de los 8 números del cromosoma por otro, por ejemplo, [23126675] → [23026675]. El cruzamiento consiste en el intercambio, a partir de cualquier posición, de la información de los cromosomas de los individuos seleccionados. Por ejemplo, [12345678] + [87654321] → [12354321].

Con estas definiciones, el algoritmo genético está adaptado al problema y su ejecución consiste en elegir una población inicial de N individuos (x_i, y_i), seleccionarlos según su adaptación usando la función $f(x_i, y_i)$ y aplicarles los operadores genéticos para generar la nueva población (Ison et al., 2005).

En el presente estudio, el *k*- NN en combinación con GAs se ejecutó en varias ocasiones, las necesarias para generar modelos de clasificación para cada uno de los índices bióticos obtenidos y sus clases asignadas *a priori*; se efectuaron los respectivos análisis comparativos entre ellos y con esto se pudo determinar cual índice es el que presenta un mejor ajuste matemático para un modelo de clasificación. Con este planteamiento, también se logró verificar cuáles de las variables descriptoras tienen una mayor influencia en la correcta asignación de objetos a las clases de calidad biótica.

6.5. El *k* – NN y su aplicación frente a la primera pregunta científica

El primer objetivo a lograr a través de la gran y compleja matriz de datos de calidad del agua, fue responder a la pregunta ¿Cuál índice biótico dado por la comunidad de macrozoobentos es el óptimo como medida de respuesta biológica para un modelo de clasificación? Tal es así, que de una lista de métricas bióticas (que se detallan más adelante) se construyeron modelos de clasificación *k* – NN + GAs para cada una de ellas. Estas medidas bióticas determinaron las clases (por ejemplo 1, 2, 3) a las cuales se pretendió que las estaciones de muestreo, también clasificadas en 1, 2, 3 (según el índice biótico), y descritas únicamente por variables descriptoras (físico – químicas, microbiológicas y geomorfológicas) sean asignadas. Lógicamente, para un modelo óptimo se supuso un mayor número de estaciones correctamente asignadas, es decir, por ejemplo, que puntos de monitoreo de la clase 3, se asignen a la clase que en teoría le corresponde, la 3. El índice biótico que presentó mayor número de estaciones de muestreo correctamente asignadas fue el elegido como óptimo para la CRP (mayor NER %). Se entiende que la 'asignación correcta de patrones' está directa y proporcionalmente relacionada con la capacidad del índice biótico de reflejar lo que los descriptores físico – químicos, microbiológicos y geomorfológicos, condicionan. Es decir, mientras mayor sea esta capacidad será una mejor 'variable de respuesta biológica'.

6.6. El *k* – NN; GAs y la calidad de las aguas superficiales (nivel regional y mundial)

A nivel regional en Ecuador se cuenta solamente con un reporte bibliográfico referente al empleo del *k* – NN en temas ligados al agua. Sotomayor (2007) utilizó con éxito el *k* – NN para evaluar los efectos de la actividad minera en la microcuenca del río Tenguel, ubicada en la parte sur de la costa de Ecuador. En este trabajo, el algoritmo de clasificación empleado sirvió

para definir grupos de estaciones de muestreo según su grado de contaminación así como también las variables latentes que más explicaban a estos grupos.

En un contexto medianamente similar, un aporte importante para Ecuador estuvo dado por Domínguez-Granda et al. (2011). En este trabajo, se utilizaron árboles de agrupación multi - objetivo con el fin de predecir índices de buen rendimiento ecológico para calidad del agua, y al mismo tiempo, la ocurrencia de taxones de macroinvertebrados claves, todo esto a la par y sobre la base de las variables ambientales más importantes. Este estudio se realizó en la cuenca del río Chaguana al sur de la costa del país, y puso de manifiesto que los árboles de agrupación multi - objetivo se pueden utilizar fácilmente como una herramienta práctica para apoyar a las decisiones de los gestores de calidad del agua en el país.

A nivel mundial, el estudio del estado del arte sugiere que el potencial del *k* – NN ha sido pobremente explorado en el contexto de los recursos hídricos pues tan solo se ha logrado identificar una sola publicación. Librando (1991) llevó a cabo un trabajo con el fin de reducir el volumen de información necesaria para la evaluación de la calidad de algunas aguas superficiales designadas para la purificación con datos recogidos en un estudio previo en tres ríos (Simeto, Alcántara y Oreto) en Sicilia. En este trabajo se efectuaron varios análisis estadísticos multivariados y entre algunos de los métodos empleados estuvo el *k* – NN, el cual brindó interesantes resultados pues las variables medidas se agruparon en diferentes conglomerados según su contenido de información. Esto proporcionó un criterio fiable para la elección de los parámetros óptimos.

Alrededor del mundo otras técnicas similares al *k* – NN son mayoritariamente empleadas con el fin de evaluar las grandes y complejas matrices de datos de calidad de agua superficial. Métodos de clasificación más robustos (paramétricos) como el Análisis Discriminante (DA) han sido utilizados con éxito en países tan diversos como Argentina (Wunderlin et al., 2001); Polonia (Kowalkowski et al., 2006); Nepal (Kannel et al., 2007); Japón (Shrestha & Kazama, 2007); Malaysia (Juahir et al., 2009); Turquía (Koklu et al., 2010); etc., Básicamente en estos estudios lo que se ha pretendido siempre es una reducción del espacio multidimensional para una interpretación más cómoda de la variabilidad de la calidad de agua, tanto a nivel temporal como espacial.

Referente a los GAs, D'heygere et al. (2003) estudiaron una base de datos de mediciones recogidas en 360 puntos de muestreo en los cursos de agua no navegables en Flanders (Bélgica), con el fin de predecir la presencia / ausencia de taxones de macroinvertebrados bentónicos mediante árboles de decisión. El poder predictivo del método fue analizado en base al porcentaje de objetos correctamente clasificados (el mismo contexto que el considerado en este estudio) y un algoritmo genético se introdujo para comparar el poder predictivo de los diferentes conjuntos de variables de entrada en los árboles de decisión. Como resultando se obtuvo que el número de variables de entrada se redujo de 15 a 2 - 8 variables sin afectar el poder predictivo de los árboles de decisión de manera significativa.

Un trabajo afín se llevó a cabo en Etiopia (Ambelu et al., 2010). Muestras de macroinvertebrados bentónicos y datos de parámetros físico - químicos fueron obtenidos en la cuenca del río Gilgel Gibe en el sudoeste de ese país durante el período 2005 – 2008. En una siguiente etapa, las métricas ecológicas calculadas se compararon entre sí para evaluar su relevancia (contexto similar que el considerado en este estudio). Árboles de Clasificación (CTs) y las Máquinas de Soporte Vectorial (SVMs) se utilizaron para inducir modelos que describan la relación entre las características de los ríos y las condiciones ecológicas. Luego, los GAs se aplicaron con el fin de mejorar el rendimiento de los modelos construidos y para la fácil interpretación de estos al hacer una selección de las variables de entrada que más se utilizaron.

Asimismo, Hoang et al., (2010) evaluaron los CTs y las SVMs con el propósito de estudiar la idoneidad del hábitat para 30 taxones de macrozoobentos en el río Du al norte de Vietnam. La presencia / ausencia de estos 30 taxones se modeló en base a 21 variables de tipo físico – químicas. El rendimiento estadístico de ambos métodos multivariados fue analizado en base al porcentaje de objetos correctamente clasificados (el mismo enfoque que el considerado en este estudio), y el Coeficiente kappa de Cohen, resultando que la capacidad predictiva de SVMs fue superior. Los autores concluyen que SVMs ha demostrado tener un alto potencial cuando se aplica en la toma de decisiones en el contexto de la restauración de ríos y la gestión de los mismos.

En sí, a nivel mundial hay una creciente tendencia a la modelización de varios tipos de ecosistemas y a la predicción de los organismos de agua dulce a base de técnicas de aprendizaje automático que se están volviendo más y más fiables debido a (1) la creciente disponibilidad de grandes bases de datos; (2) las técnicas avanzadas de modelización; y (3) la progresiva potencia de cálculo de los actuales ordenadores. Pero en sí, la aplicación de análisis como los GAs en el contexto del adecuado manejo y gestión de los recursos hídricos tienen un inicio relativamente reciente, de tal modo que no hay un número grande de trabajos publicados. Sin embargo, el Laboratory of Environmental Toxicology and Aquatic Ecology de la Ghent University (Flanders, Bélgica) ha efectuado numerosos trabajos en el ámbito discutido.

6.7. Análisis de Componentes Principales

El análisis de componentes principales (PCA) de muchas formas constituye la base de los análisis de datos multivariados (Wold & Geladi, 1987). Es una potente técnica multivariante propuesta a principios del siglo 20 y aplicada por primera vez en ecología a mediados del mismo con el nombre de análisis de factores (McCune & Grace, 2002). Es de fundamental importancia en la exploración de datos, y en síntesis y de manera general, en el PCA las variables que describen los datos se transforman en nuevas variables denominadas, componentes (factores o variables latentes). Estas nuevas variables son combinaciones lineales estandarizadas de las variables originales y son ortogonales entre sí (Ouyang, 2005).

De tal modo, el punto central del PCA es reducir la matriz original (m, n) de datos X compuesta de m variables que describen o caracterizan a n objetos (puntos de muestreo en el presente estudio), a saber:

- Pesos $\boldsymbol{a}$
- Scores $\boldsymbol{f}$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ a_{21} & \cdots & a_{2s} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{ms} \end{pmatrix} \cdot \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ f_{s1} & f_{s2} & \cdots & f_{sn} \end{pmatrix}$$

$$X = a * f$$

donde, X = Matriz de datos originales

a = Pesos,

f = Scores, y

s = Número de factores $(s = m)$

Una combinación lineal de los diferentes factores de la matriz $\boldsymbol{a}$ con los Scores de $\boldsymbol{f}$ pueden reproducir la matriz de datos $\boldsymbol{X}$. Estos nuevos factores son variables sintéticas y representan una cierta cantidad de información del conjunto de datos; explican el total de la variabilidad de todas las características o variables medidas en un orden descendente y no se encuentran correlacionadas entre sí (Zwanziger et al., 1997). Es, por lo tanto, posible reducir la dimensionalidad de m para los datos con un mínimo de pérdida de información expresada por la matriz de residuos $\boldsymbol{e}$.

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ a_{21} & \cdots & a_{2s} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{ms} \end{pmatrix} \cdot \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{s1} & f_{s2} & \cdots & f_{sn} \end{pmatrix}$$

$$+ \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{pmatrix}$$

$$X = a * f + e$$

donde, X = Matriz de datos originales

e = Residuos como resultado de la reducción de la dimensionalidad; y

s = Número de factores $(s < m)$

De esta manera, el PCA (Fig. 6) provee una proyección de los objetos desde un espacio caracterizado por una alta dimensionalidad hacia un espacio definido por pocos factores (Zwanziger et al., 1997).

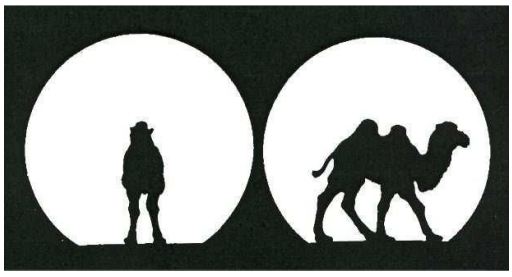


Figura 6. La idea del PCA, encontrar la proyección correcta (Gonze, 2007).

Algunos autores como Kendall, (1980) y Cadima & Jolliffe, (1995) creen que la interpretación de los nuevos ejes de ordenación posee un componente de subjetividad. Esto es debido, en parte, al hecho de que las herramientas de ordenación se utilizan principalmente como medios de exploración de los datos en lugar de que se focalicen en comprobar o rechazar una hipótesis para una prueba dada. Con ello, sea cual fuere el objetivo del PCA (exploración inicial o aceptar o rechazar una hipótesis) a fin de proporcionar una significativa y correcta interpretación de los nuevos componentes principales, es importante determinar qué variables (de las originales) están asociadas mayoritariamente a estos, pues la capacidad de identificar relaciones y minimizar el efecto de variación aleatoria puede contribuir sustancialmente al reconocimiento de patrones significativos en los datos. De tal modo, los resultados de ordenación en el PCA se interpretan habitualmente en la siguiente forma (Peres-Neto et al., 2003):

- A través de valores propios resumidos por los ejes de ordenación (nuevos PC) que cuantifican la cantidad de variación de los datos originales.
- Vectores propios que contienen los coeficientes que relacionan las variables originales a los nuevos ejes de ordenación. Estos se conocen con el nombre de 'Loadings' (Pesos).
    - Alguno autores interpretan los coeficientes de los vectores propios escogiendo de forma arbitraria un valor (ej., 0,5) para así señalar su importancia o insignificancia.

En contexto, el PCA se puede llevar a cabo a nivel de grupos de datos predeterminados en gabinete. El análisis se desarrolla entre las medias de los grupos (es decir, los elementos analizados son los grupos y no los objetos como tales) y los puntajes de PCA se calculan utilizando el vector de productos con los datos originales (Reisenhofer et al. 1998; Hammer, Harper & Ryan, 2007).

6.8. El PCA y su aplicación frente a la segunda pregunta científica

Como se explicó, el PCA se puede llevar a cabo a nivel de grupos de datos predeterminados en gabinete; en el presente estudio, estos grupos se condicionaron por el índice biótico que fue elegido como óptimo al responder la primera pregunta científica, de esta forma se cumplió el criterio de respuesta biológica que exige la EH, pues el PCA se desarrolló entre los grupos de estaciones de muestreo, y estas estuvieron descritas únicamente por variables físico – químicas, microbiológicas y geomorfológicas, pero condicionadas directamente por el índice biótico.

Asimismo con esta metodología se restó mucho peso a la discontinuidad existente en los datos disponibles para el presente estudio, puesto que el PCA no se realizó sobre estaciones cuya representación en el muestreo total fue muy baja o muy alta, sino sobre grupos o clases de calidad biótica del agua.

6.9. El PCA y la calidad de las aguas superficiales (nivel regional y mundial)

Al ser el PCA una técnica con más de cien años de creación, su diseminación en diferentes campos de las ciencias exactas y naturales ha sido amplísima. En el manejo y gestión de recursos hídricos superficiales su uso ha consistido principalmente en la reducción del espacio multivariante de datos, en la caracterización de cambios a nivel espacial y temporal de parámetros de calidad de agua y en la identificación de las principales variables latentes (Ouyang, 2005).

A nivel regional en Ecuador un trabajo interesante que vincula a la calidad de agua de un sistema hídrico (río Chaguana) y el uso del PCA (Domínguez-Granda et al., 2005) tuvo lugar con el fin de explorar las correlaciones entre macroinvertebrados acuáticos y distintas variables de tipo físico – químicas, pesticidas, así como varios usos del suelo y cotas altitudinales. Las estaciones muestreadas, al ser caracterizadas por sus variables ambientales, mostraron un gradiente en relación a su altitud, principalmente estructurado por la concentración de amonio en el sedimento, el ancho del río, la conductividad y la temperatura del agua. Dicho gradiente no mostró patrón alguno al ser evaluado en relación al orden del río, en tanto que lo hizo parcialmente en relación al uso del suelo.

Otro importante aporte se llevó a cabo con el fin de evaluar los efectos de la deforestación sobre la comunidad de peces en la amazonía ecuatoriana (Bojsen & Barriga, 2002). En este caso, el PCA realizado sobre mediciones en la estructura de la comunidad de peces mostró que seis variables ambientales (la zona de fondo del arroyo cubierto de hojas, el área de piscina o poza, partículas de materia orgánica, la profundidad, la conductividad y la media de sólidos

suspendidos) estaban relacionadas con los ejes de ordenación. La presencia de hojas, que se correlaciona fuertemente con la cubierta de dosel, fue la variable más estrechamente relacionada con la estructura de la comunidad de peces, mientras que el área de la piscina relativa fue la variable que secundaba en importancia. Los autores concluyeron que la estructura de la comunidad de peces se vio fuertemente afectada por la deforestación.

En un ámbito y contexto similares, Gallegos (2013) produjo un valioso aporte al estudiar a través de un método de ordenamiento afín al PCA como es el Análisis de Correspondencias Canónicas (CCA), el efecto de la cobertura de la vegetación ribereña, la estructura trófica y la estacionalidad climática sobre la comunidad de macrozoobentos en los andes ecuatorianos (río Sambache, Refugio de Vida Silvestre Pasochoa). El CCA mostró que la conductividad eléctrica y la cantidad de sólidos disueltos totales fueron los parámetros más importantes para explicar la estructura de ensamblaje de los macrozoobentos.

A nivel mundial existen un sin número de artículos en los cuales el PCA ha sido utilizado como método de exploración de datos de calidad de agua superficial. Su uso ha tenido una dispersión en todo el globo. Así, De Ceballos et al. (1998) en un estudio llevado a cabo para tres lagos localizados al noreste de Brasil determinó a través del PCA que las clases de calidad de agua estuvieron definidas por 9 variables, agrupadas en dos componentes principales.

Vega et al., (1998) aplicó el PCA sobre datos levantados en el río Pisuerga en la cuenca del Duero (Centro Norte de España) identificando un número reducido de factores latentes como fueron los contenidos minerales, elementos antrópicos y la temperatura del agua.

En Nigeria en la cuenca del río Jakara, ubicada al noreste del país, Mustapha & Abdu (2012) emplearon el PCA con el fin de determinar fuentes de polución y la contribución de estas en la variación espacio temporal de la calidad del agua. El PCA atinó que las variables más importantes que determinan la calidad del agua en el sistema hídrico fueron la erosión, los desechos domésticos, los efectos de dilución y la escorrentía de zonas agrícolas.

Otro trabajo interesante mediante el PCA fue llevado a cabo en el río St. Johns en Florida, E.E.U.U. (Ouyang, 2005), aunque en este caso el análisis fue dado con el fin de evaluar la importancia de las estaciones de muestreo de la red que se tenía montada en ese cuerpo de agua. Los resultados mostraron que tres sitios de muestreo fueron clasificados como menos importantes para explicar la variabilidad anual del conjunto de datos de calidad de agua y por tanto podían ser excluidos de la red.

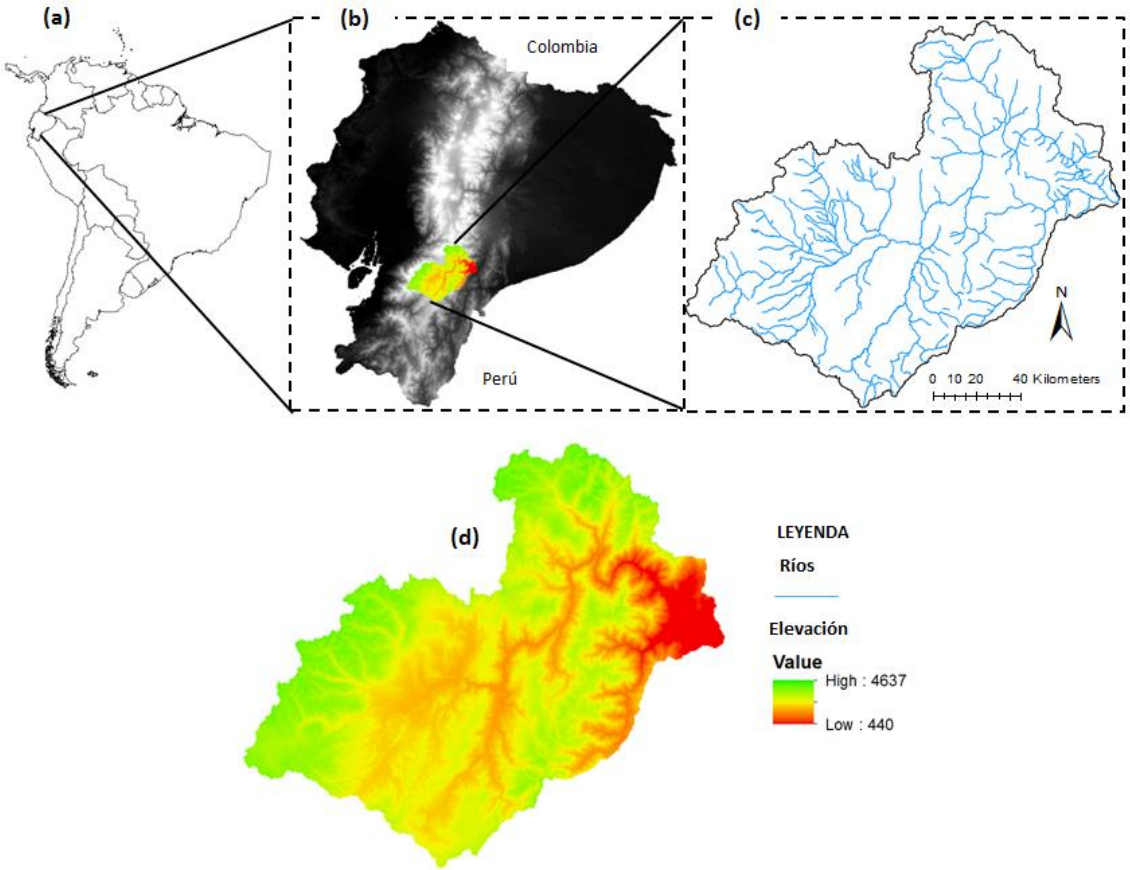
También son importantes ejemplos de los usos del PCA ligados a los contextos de calidad y ecología acuática. López-Luna et al. (2012) efectuaron experimentos con *Oreochromis niloticus* y sus hábitos alimenticios y la influencia de estos sobre la calidad de agua. Ellos sugieren que variables como el contenido total de iones disueltos y los hábitos alimenticios del pez son las variables más importantes para describir la variabilidad del sistema analizado.

A saber, un sin número de trabajos que involucran al PCA y variables ligadas a la temática de los recursos hídricos superficiales, su manejo y gestión adecuadas, se publican y se siguen generando en la actualidad.

7. MATERIALES Y MÉTODOS

7.1. Zona de estudio

La CRP cuenta con una extensión de 6439 km^2 y una población de unos 900000 habitantes. Su extensión se reparte en tres provincias (Azuay, Cañar y Morona Santiago); y se ubica en la zona centro sur del Ecuador, entre las latitudes -2,30° y -3,27° y longitudes -78,26° y -79,36° (Mapa 2). Es la cuenca hidrográfica más importante del Ecuador dada la explotación de su enorme potencial hidroeléctrico[4] pues permite abastecer en la actualidad más del 40% de la demanda energética del país. Este sistema hídrico forma parte de las cuencas interandinas centrales del Ecuador y su orientación es de Suroeste a Noreste. El río Paute desemboca en el río Upano, formando parte del sistema hidrográfico amazónico que desagua en el océano Atlántico. Un detalle interesante para la CRP es dada la subducción de la placa de Nazca hacia la placa de Sudamérica, lo que causa que esta cuenca hidrográfica siga en proceso de levantamiento, agudizando su pendiente, lo cual acarrea grandes cantidades de sedimentos en suspensión hacia la Amazonia (Astudillo et al., 2010).



Mapa 2: Ubicación de la CRP en el contexto de (a) América del Sur; y de (b) Ecuador; (c) propiedades hidrográficas y (d) topográficas (MDT).

[4] A nivel nacional, la cuenca tiene una mayor importancia en la generación de energía hidroeléctrica. La empresa estatal Corporación Eléctrica del Ecuador (CELEC-EP), a través de la Unidad de Negocio Hidropaute, tiene a su cargo la operación de las Centrales Mazar y Molino, como parte del Proyecto Paute Integral, cuatro centrales en cascada que aprovechan el agua de la CRP. La Unidad de Negocio Hidropaute cuenta con un potencial de generación eléctrica instalado de 1.100 MW en la Central Molino y de 170 MW en la Central Mazar. Entre el año 2000 y 2010, la Central Molino tuvo una producción energética anual que oscila entre 4.049 GWh y 6.286 GWh, dependiendo de las variaciones anuales en precipitación.

Otro importante factor a considerar para la CRP radica en que ella se ubican áreas protegidas y ecosistemas únicos, entre los que han de destacarse el Parque Nacional "El Cajas" (PNC) (zona occidental) el cual es un humedal de importancia internacional RAMSAR y el Parque Nacional "Sangay" (zona nororiental), ambos reconocidos por la UNESCO como Patrimonio Natural de la Humanidad.

Asimismo, históricamente la CRP ha sido considerada como una cuenca piloto estratégica para la promoción de una Gestión Integrada de los Recursos Hídricos (GIRH), una filosofía basada en la mediación de conflictos que ha encontrado en estos últimos años no pocas dificultades prácticas para su implantación en el Ecuador, como consecuencia de la compleja realidad institucional, social, étnicocultural y económica del país. De este modo, en esta cuenca se constituyó el primer Organismo de Cuenca del Ecuador, mediante ley publicada en el Registro Oficial el 9 de noviembre de 2005, denominado Consejo de Aguas de la Cuenca del Paute (CG Paute) y configurado como un organismo público y descentralizado, que integraba a los gobiernos provinciales y municipales, entidades del sector público (gobierno central) y privado, usuarios y otros actores clave implicados en la gestión de la cuenca (Molina, 2008).

En contexto, la CRP al igual que muchos sistemas hídricos de Ecuador es muy heterogénea por naturaleza en casi todos sus aspectos. Forma parte del callejón interandino, con pendientes que se encuentran en el rango de 25 a 50%; el relieve escarpado es representativo de la zona media y baja, y le sigue un relieve montañoso. Los rangos altitudinales varían entre los 500 y 4250 m.s.n.m. En lo referente al clima, en general en Ecuador este varía de caluroso a frío; los Andes que atraviesan la geografía del país han generado diferentes regiones climáticas. Así, las zonas andinas son muy frías, con temperaturas menores a los 10 °C; mientras que en las regiones del litoral y de la selva, las temperaturas promedio superan los 30 °C. La temperatura de la CRP cuenta con un rango de media multianual que va desde los 4.4 a los 18.6 °C. Las Zonas de menor temperatura corresponden a la cumbres de la cordillera Occidental de los Andes con un promedio de temperatura media de 6 °C (Páramo), en tanto las zonas más cálidas se encuentran en los valles interandinos antes descritos y en el oriental de la zona del subtrópico hacia la Amazonía, con promedios de 22 a 26 °C.

Igualmente dado al amplio rango altitudinal, su régimen de precipitaciones es muy variado en intensidad y duración, teniendo promedios máximos anuales de 2500 – 3000 mm en el extremo oriental de la cuenca y también la ocurrencia de precipitaciones máximas de 1200 a 1500 mm en la línea de cumbres de la Cordillera Occidental en la franja occidental de la cuenca. A su vez los promedios mínimos anuales de precipitaciones están entre los 600 y los 800 mm y se registran en los valles interandinos.

Todas estas características a nivel biofísico y en lo que se refiere estrictamente a la productividad hidroeléctrica de la zona, hacen de la CRP un sistema hídrico muy especial y determinante para el país. Dada la importancia de la zona, en la CRP un total de 64 puntos de muestreo fueron evaluados en varios trabajos efectuados desde el año 2008 hasta el 2010 – 2013.

En este contexto, la matriz producto de la sistematización de la información de calidad de agua es el insumo macro de este estudio y los trabajos que contribuyeron para construirla fueron:

1. Estudio de calidad de agua en la Provincia del Azuay, realizado por el CREA, UDA (Universidad del Azuay) y el IRHA. Muestreos mensuales en enero, febrero y marzo de 2008.

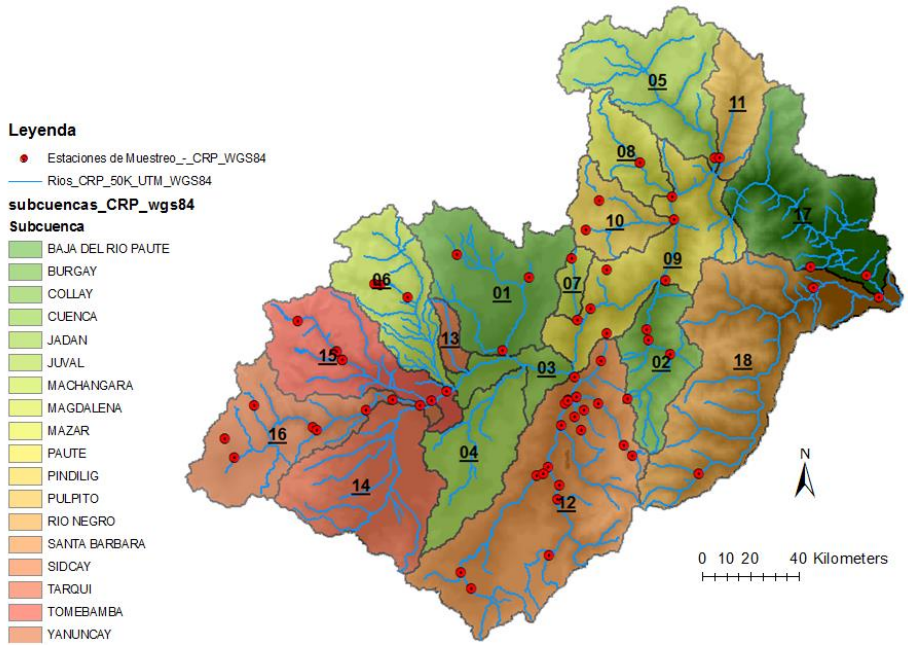
2. Monitoreo de la calidad de agua en la CRP, 10 Subcuencas. Autor: UDA, bajo contrato con el ex CG Paute. Muestreos bimensuales desde octubre de 2010 hasta abril de 2011.

3. Monitoreo de la calidad de agua en la CRP, 10 Subcuencas. Autor: UDA, bajo contrato con el ex CG Paute. Muestreos bimensuales desde octubre de 2011 hasta abril de 2012.

4. Monitoreo de la calidad de agua en la CRP, 10 Subcuencas. Autor: UDA, bajo convenio con la SENAGUA - DHS. Muestreos bimensuales desde agosto de 2012 hasta marzo de 2013.

7.2. Variables evaluadas

Varios tipos de variables fueron levantadas durante los programas de monitoreo elaborados en la CRP, estas son del tipo físico – químicas, microbiológicas, biológicas (representadas por los macrozoobentos identificados a nivel de familia) y también valoraciones de la calidad del hábitat (bosques riparios y calidad del lecho). Todas formaron parte de la matriz objeto del presente análisis y a continuación se las describe detalladamente. A saber, adicionalmente ciertos parámetros geomorfológicos fueron calculados para el presente estudio de caso.



Mapa 3. Red de las 64 estaciones de muestreo para el monitoreo de la calidad de agua en la CRP. Las subcuencas de la CRP y su respectivo código numérico identificador se corresponden con los datos de la Tabla 1.

7.2.1. Físico – Químicas y Microbiológicas

Un total de 27 variables físico químicas y 2 microbiológicas (*) fueron analizadas en los 64 puntos de muestreo a lo largo de la serie temporal considerada (Tabla 2). Las metodologías empleadas para los análisis fueron las desarrolladas por los "Standard Methods" de "USA" y el ente ejecutor de dichas marchas metodológicas fue el Laboratorio de Química Ambiental y Microbiología de la Universidad del Azuay (UDA), Cuenca – Ecuador.

Tabla 1. Georeferenciación de las 64 estaciones de muestreo (UTM) y datos de la subcuenca a la que pertenecen.

Estación (Hydrocode)	X - ME	Y - MN	Altitud (m.s.n.m.)	Subcuenca	Código - Subcuenca	Área / Subcuenca (ha)
B13	740378	9699833	2560	Burgay	01	44703
B2	735897	9687554	2320			
B8	728099	9703754	2980			
COL1	757017	9679274	3400	Collay	02	23936
COL2	764289	9686904	2600			
COL3	760682	9689363	2240			
COL4	760272	9691084	2220			
J1	771865	9720193	2040	Juval	05	42732
P1	772670	9720229	2160	Púlpito	11	16921
MAC9	715134	9698594	3360	Machángara	06	32545
MAC10	719701	9696476	3160			
MAC6	714156	9698843	3520			
MAG1	747648	9703135	3000	Magdalena	07	5081
MAG2	748549	9692642	2290			
MAZ1	764614	9713596	2020	Mazar	08	16577
MAZ3	759152	9719434	2520			
MAZ4	759135	9719317	2520			
N1	799668	9696494	480	Negro	18	80222
N2	788624	9698144	800			
N3	769119	9666642	2120			
PAU6	763469	9699364	2420	Paute	09	44712
PAU8	750845	9694527	2160			
PAU11	753550	9701168	2920			
PAU12	753467	9690378	2960			
PB2	797652	9700177	480	Parte Baja del Paute	17	51013
PB3	788119	9701700	1600			
PIN1	764932	9709722	2060	Pindilig	10	16827
PIN4	752175	9712862	2920			
PIN5	750001	9708006	3360			
O58	748060	9682960	2220	Santa Bárbara	12	95252
O60	746923	9679101	2240			
O61	757878	9669696	3180			
O62	756397	9671419	3180			
O63	752160	9678481	2560			
O64	748439	9679687	2240			
O65	746510	9678516	2240			
O66	748133	9676303	2500			
O67	749243	9674093	2660			
O68	749621	9677445	2460			
O69	746910	9678992	2240			
SB1	745516	9664770	2420			
SB2	745890	9674898	2260			
SB3	752566	9685756	2780			
SB4	743685	9652806	2760			
SB5	728713	9649895	2640			
SB6	730515	9647122	2680			
SB7	741581	9666327	2340			
SB8	745106	9662310	2400			
SB9	743631	9667803	2300			
SB11	742729	9666655	2320			
TOM2	723861	9679008	2480	Tomebamba	15	38041
TOM8	701131	9692561	3640			
TOM11	708644	9685960	3040			
TOM12	707634	9687398	3200			
TOM14	708645	9685936	3040			
TOM18	726356	9680533	2450			
YAN1	721781	9678176	2510	Yanuncay	16	41888
YAN2	717206	9679147	2610			
YAN3	690320	9669339	3580			
YAN4	693634	9678203	3780			
YAN5	688716	9672508	3660			
YAN6	712631	9677344	2750			
YAN7	703622	9674467	3000			
YAN8	704286	9673958	2980			
Estas subcuencas no fueron monitoreadas				Cuenca	03	12029
				Jadán	04	29751
				Sidcay	13	4330
				Tarqui	14	47629

Tabla 2. Variables físico – químicas analizadas. (*) Microbiológicas.

Parámetro Medido	Unidad	Símbolo / Abreviatura	Límite de Detección	Media	Desviación Estándar
Calcio	(mg/L)	Ca	0,15	4,06	5,47
Nitratos	(mg/L)	NO_3	0,01	0,56	2,05
Amonio	(mg/L)	NH_4	0,02	0,78	1,57
Dureza Total	(mg/L)	$CaCO_3$	-	33,34	36,50
Alcalinidad	(meq/L)	-	-	0,74	1,35
Turbiedad	(UNT)	UNT	-	19,48	88,16
Demanda Bioquímica de Oxígeno	(mg/L)	DBO	-	10,30	8,08
Demanda Química de Oxígeno	(mg/L)	DQO	-	24,44	17,56
Sólidos Totales	(mg/L)	TS	-	2,00	10,42
Sodio	(mg/L)	Na	0,02	5,28	9,15
Hierro	(mg/L)	Fe	0,021	0,21	0,49
Cadmio	(mg/L)	Cd	0,21	0,01	0,05
Magnesio	(mg/L)	Mg	0,12	2,23	3,97
Potasio	(mg/L)	K	0,02	1,63	5,29
Plomo	(mg/L)	Pb	0,2	0,02	0,07
Cloruros	(ug/L)	CL-	0,18	5,26	27,32
Floruros	(ug/L)	F-	0,2	1,56	6,83
Fosforo	(mg/L)	P	0,35	0,38	0,50
Fosfatos	(mg/L)	PO_4	0,93	1,29	1,58
Cobre	(mg/L)	Cu	0,46	0,02	0,12
Níquel	(mg/L)	Ni	0,2	0,02	0,14
Aluminio	(mg/L)	Al	0,01	0,06	0,23
Potencial de Hidrógeno	-	pH	-	7,52	0,65
Temperatura	(°C)	T(°C)	-	14,57	3,27
Conductividad	(µs/cm)	EC	-	122,44	197,36
Oxígeno Disuelto	(mg/L)	OD	-	6,83	0,75
Saturación de Oxígeno	(%)	% SAT O_2	-	89,82	6,54
Coliformes Totales*	(NMP/100ml)	Col - T	-	7045,74	7035,34
Coliformes Fecales*	(NMP/100ml)	Col - F	-	5038,32	6451,66

La toma de muestras se realizó procurando que el muestreo sea lo más confiable posible, así los puntos de colección fueron representativos del tipo de hábitat de estudio. Se seleccionaron zonas donde el agua estuvo bien mezclada (zonas centrales), evitando tomar colectas superficiales, rebosaderos de los embalses, confluencias de ríos poco importantes, lugares de pequeños vertidos, etc., ya que éstos sólo tienen efectos muy localizados en la química del agua de ese tramo. Las colectas de agua realizadas (para análisis físico – químicos y microbiológicos) se conservaron en frascos limpios y a temperaturas constantes entre 3 y 8 °C dentro de una nevera con bloques helados hasta su posterior análisis en laboratorio (tiempo máximo de transporte hasta el laboratorio 3 ± 1.5 horas). Esto, siguiendo la metodología del protocolo GUADALMED (Protocolo Rápido de Evaluación de la Calidad Ecológica ó PRECE) diseñado por diferentes universidades de España (Jáimez-Cuéllar et al., 2002).

7.2.2. Calidad de Hábitat

La valoración de la calidad del hábitat es fundamental en cualquier evaluación de la integridad ecohidrológica en los sistemas de agua lóticos, y debe realizarse en cada sitio en el momento de la toma de muestras (Barbour et al., 1999). Esto, debido a que el hábitat y la diversidad biológica en los ríos están estrechamente ligados (Raven et al., 1998). El 'hábitat' incorpora todos los aspectos tales como los componentes físicos y químicos, junto con las interacciones bióticas (Barbour et al., 1999) y normalmente su definición es reducida a la calidad del lecho del río y la vegetación riparia. La presencia de alteraciones en la estructura del hábitat es considerada como un fuerte factor de estrés en los sistemas acuáticos (Karr et al., 1986). De tal modo, las razones para evaluar la calidad de riberas de los ríos se justifican dados sus valores naturales como elevada riqueza, diversidad florística y faunística (Girel & Manneville, 1998; Ward, 1998), a su capacidad para incidir sobre la calidad ambiental del ecosistema acuático que rodea a través del control de la temperatura del agua (Beschta et al., 1997) y por el aporte de materiales orgánicos externos al lecho o canal (Fisher & Linens, 1973) y de los nutrientes

(Schade et al., 2001, 2002). Además, la vegetación de ribera juega un papel esencial en la retención y atenuación de los efectos destructores de las avenidas de agua (Decamps, 1996). Como zona de transición o interface ha sido muy estudiada su función de filtro y en cierta medida de sistema depurador (Osborne & Kovacic, 1993). Todos estos valores y sobre todo funciones que desempeñan los bosques de ribera les hacen excelentes indicadores del uso que se da al suelo, y por ende, si este condiciona positiva o negativamente a la biodiversidad acuática y a la calidad del agua en general (Fig. 7).

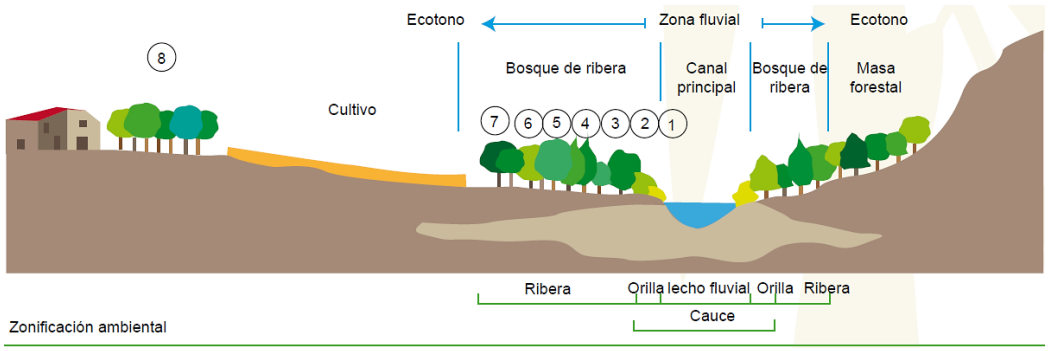


Figura 7. Orillas o márgenes de río (ecosistemas ribereños; Camprodon et al., 2012).

En contexto, la calidad del lecho es directamente relacionada con la capacidad del hábitat físico para albergar una fauna determinada, entendiéndose que a mayor heterogeneidad y diversidad de estructuras físicas del hábitat, le corresponde una mayor diversidad de comunidades biológicas que lo colonizan (Suárez et al., 2002). La heterogeneidad del lecho se considera en la actualidad como uno de los principales factores de influencia de la riqueza de especies de invertebrados acuáticos (Voelz & MacArthur, 2000). Al estudiarla se procura diagnosticar aspectos físicos del cauce como la heterogeneidad de hábitat y que dependen en gran medida de la hidrología y del sustrato existente, entre ellos la frecuencia de rápidos, los regímenes de velocidad y profundidad, el grado de inclusión y sedimentación en pozas y la diversidad de sustratos (Fig. 8).

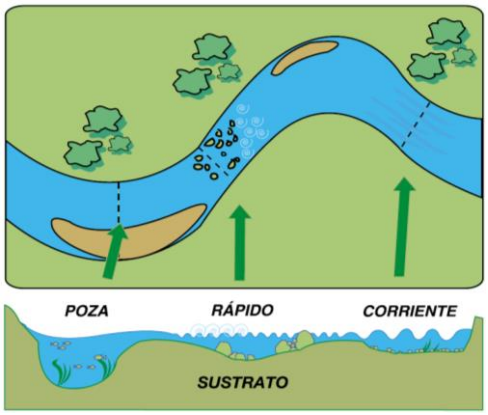


Figura 8. Esquema del lecho de un río (Carrera & Fierro, 2001).

La evaluación del hábitat se define como la valoración de las estructuras que rodean físicamente al cuerpo de agua y que influyen en la calidad del mismo y en el estado de las comunidades acuáticas residentes de este (Barbour et al., 1996). Para los ríos, un enfoque que abarque la evaluación de la estructura del hábitat incluye una evaluación de la variedad y la calidad del sustrato, la morfología del canal, la estructura del banco, y vegetación ribereña.

Rosgen (1985, 1994) identificó y definió ocho variables que determinan la estabilidad de la morfología del canal o lecho, las cuales no son independientes entre sí, estas son:

- La anchura del cauce.
- Profundidad del canal.
- La velocidad del flujo.
- El caudal.
- La pendiente del canal.
- La rugosidad de los materiales de canal.
- La carga de sedimentos.
- La distribución del tamaño de las partículas.

Cuando los ríos tienen una de estas características alteradas, su capacidad para disipar energía correctamente se pierde y esto se traduce en el aumento de las tasas de erosión del canal (Leopold et al., 2012; Rosgen, 1985). Gran parte de la comprensión de las relaciones del hábitat en los ríos ha surgido de los estudios comparativos que describen las relaciones estadísticas entre variables del hábitat y la abundancia de la biota (Hawkins et al., 1993). Sin embargo, en respuesta a la necesidad de incorporar las evaluaciones de amplia escala del hábitat en los programas de monitoreo de recursos hídricos, han surgido muchas alternativas (protocolos). De tal forma, para el presente estudio se escogió uno del tipo rápido y cualitativo para describir la calidad general del hábitat físico, (Plafkin et al., 1989; Rankin, 1995), es el de la EPA (Environmental Protection Agency), el cual básicamente es de tipo visual. Las variables medidas son evaluadas y clasificadas en una escala numérica en cada punto de muestreo. Los puntajes obtenidos aumentan cuando la calidad del hábitat se incrementa, de tal modo, un técnico experimentado y que esté bien instruido en la ecología y la zoogeografía de una región, puede reconocer la estructura del hábitat óptima (Barbour et al., 1999). Así, las evaluaciones del hábitat primero se hacen focalizándose en el lecho, seguidas por la morfología del canal, las características estructurales del banco y finalmente la vegetación riparia. El proceso de evaluación del hábitat real implica la calificación de ciertos parámetros como óptimos, subóptimos y marginales o pobres, dichas variables son:

1. Substrato; hábitats que podrían ser colonizados.
2. Partículas que rodean al substrato.
3. Velocidad y profundidad.
4. Acumulación de sedimento.
5. Estado del flujo del cauce.
6. Alteración del cauce.
7. Frecuencia de rápidos (o recodos).
8. Estabilidad de la orilla (cuenta cada orilla izquierda y derecha).
9. Protección de la vegetación riparia (cuenta cada orilla izquierda y derecha).
10. Ancho de la vegetación ribereña (cuenta cada orilla izquierda y derecha).

Al finalizar la valoración del hábitat se obtiene un valor respecto de 200 puntos el cual es el óptimo máximo.

Nota 1: La ficha de campo de la EPA y utilizada en este estudio para valorar la calidad de hábitat puede ser descargada del siguiente sitio Web:

http://water.epa.gov/scitech/monitoring/rsl/bioassessment/upload/ch_05.pdf

Nota 2: IHF – EPA son las siglas que corresponden con la variable 'calidad de hábitat' en este trabajo.

7.2.3. Índices Bióticos (a través de la comunidad de macrozoobentos)

Los macroinvertebrados bentónicos se describen como una abstracción que incluye a aquellos animales invertebrados que por su tamaño relativamente grande, son retenidos por redes de malla de entre 250 y 300 μm. La gran mayoría de los mismos (alrededor del 80%) corresponde a grupos de artrópodos y dentro de estos los insectos; y en especial sus formas larvarias son las más abundantes (Alba-Tercedor, 1996). Este grupo de organismos poseen una distribución a una escala global, y su sensibilidad a los cambios ambientales los hacen buenos indicadores de las condiciones hidroquímicas de un determinado cuerpo de agua superficial. De tal modo, un sin número de índices tanto de diversidad como bióticos para muestras de macrozoobentos se aplican y desarrollan en todo el mundo en un intento por medir la contaminación de los ríos, quebradas y sistemas lénticos (Giller & Malmqvist, 1998). Índices bióticos basados en puntajes, son uno de los métodos para el biomonitoreo más comunes utilizados por los gestores de recursos hídricos. En estos índices, se asigna un puntaje (score) a los taxones (por lo general se trabaja a nivel de familia) de acuerdo a su tolerancia a la contaminación, dando puntajes más altos a los taxones muy sensibles o intolerantes a efectos contaminantes y puntajes bajos a los organismos resistentes, y por ende adaptables a efectos de estrés.

Este tipo de índices se han desarrollado principalmente en Europa (Woodiwiss, 1964; Armitage et al., 1983), Sudáfrica (Chutter, 1972), América del Norte (Hilsenhoff, 1982) y Australia (Chessman, 1995) siendo uno de los más utilizados el BMWP (Biological Monitoring Working Party) (y sus derivaciones), que fue desarrollado para Reino Unido (Armitage et al., 1983).

En la CRP numerosos trabajos se han llevado a cabo a través del uso de macrozoobentos con el fin de establecer criterios de calidad de las aguas superficiales. Sin embargo no se había realizado un estudio en el cual se verifique o dictamine cual índice biótico o de estructura comunitaria es el más adecuado para afrontar a los macrozoobentos como bioindicadores. Asimismo, experiencias en la zona de estudio han permitido concluir que ciertos criterios de origen de franjas latitudinales templadas y que se aplican y toman como válidos en la CRP, no necesariamente son ciertos o correctos. Tal es el caso de Efemeróptera, la cual es una orden de insectos acuáticos que figura como un indicador de aguas limpias y bien oxigenadas (Williams & Feltmate, 1992), no obstante, la diversidad de familias y géneros que existen en las zonas tropicales[5] han hecho que esta amplia gama de diversidad permita una diversa escala de respuestas a efectos de contaminación, existiendo géneros como *Baetode*s, de la familia Beatidae (Fig. 9), que sesgan marcadamente una muestra pues siempre están presentes en hábitats limpios o contaminados, y por lo general son valorados, erróneamente, con puntajes (scores) elevados.

[5] Hasta el momento para Efemeróptera se han descrito cerca de 14 familias, 100 géneros y más de 450 especies para América del Sur (Domínguez et al., 2006).

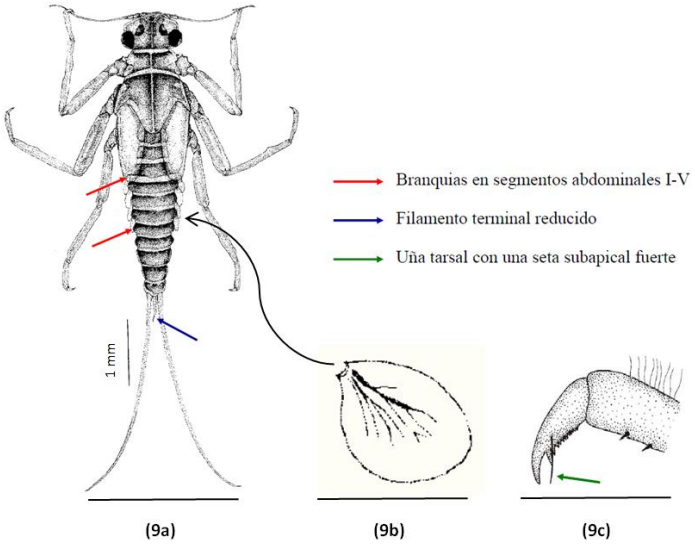


Figura 9. (a) Género *Baetodes* (apuntes de clase del Dr. Eduardo Domínguez, 2009); (b) ampliación de una agalla ventral de *Baetodes spinae* (Knox, 1964), pieza clave para su identificación taxonómica al igual que (c)**.** Cuerpo y cabeza son de color ámbar pálido, ocelo gris, antenas amarillo pálido (apuntes de clases del Dr. Eduardo Domínguez, 2009).

Con estos antecedentes, los índices que se calcularon en este estudio a través de la comunidad de macrozoobentos, y que fueron evaluados (comparados) en aras de responder el primer planteamiento científico fueron los que se indican en las siguientes subsecciones del presente documento.

7.2.3.1. BMWP (Biological Monitoring Working Party)

Desarrollado para Reino Unido por Armitage et al. (1983), este índice analiza la composición de los macrozoobentos acuáticos a nivel de familia y de acuerdo a su tolerancia a la contaminación, asignándole a cada familia un puntaje de acuerdo a su capacidad de supervivencia a distintos niveles de contaminación, 10 a los más sensibles o menos tolerantes y 1 a los tolerantes o resistentes. El puntaje final se obtiene sumando los valores de todos los componentes de cada muestra determinando así la calidad del agua (Mandaville, 2002; Tabla 1). En contexto, para este estudio, se utilizó una calibración del BMWP llamada BMWP/Col la cual fue desarrollada por Roldan (2003) para toda Colombia. De tal modo, dado que Ecuador no posee un estudio de ajuste en términos de índices bióticos y además por la cercanía y cierta semejanza con Colombia, en muchos estudios de biomonitoreo en Ecuador se usa el BMWP/Col.

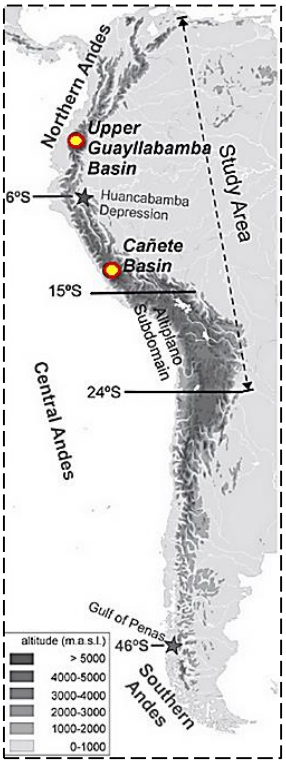
Tabla 3. Tabla para la interpretación de los valores del BMWP/Col.

Clase	Valor	Significado	Color
I	> 150, 101-120	Aguas muy limpias a limpias	
II	61-100	Aguas ligeramente contaminadas	
II	36-60	Aguas moderadamente contaminadas	
IV	16-35	Aguas muy contaminadas	
V	< 15	Aguas fuertemente contaminadas	

7.2.3.2. ABI (Andean Biotic Index)

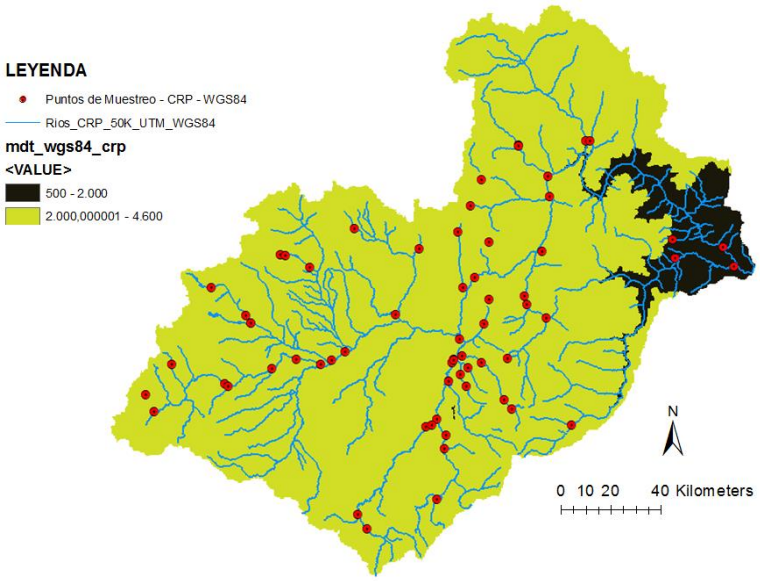
Es una extensión del BMWP presentada inicialmente en el protocolo CERA (Calidad Ecológica de Ríos Altoandinos; Acosta et al., 2009) y últimamente modificada por Ríos-Touma et al. (2014).

Está direccionado exclusivamente para las zonas de los andes (> 2000 m.s.n.m.) (Mapa 4) y en síntesis, para su calibración los autores revisaron las diferentes adaptaciones del índice BMWP que se utilizan en las regiones andinas y también examinaron los datos publicados y no publicados ("grey literature") respecto de la sensibilidad de los taxones de macroinvertebrados a la contaminación en la región. Con ello, a través de gabinete, nuevos scores fueron obtenidos para las diferentes familias de macrozoobentos.



Mapa 4. Cordillera Andina. Latitud 24° S marca el límite sur de la investigación bibliográfica de Ríos - Touma et al. (2014), y el límite norte está al final de los Andes en Venezuela. Latitud 6° S marca la zona de la depresión de Huancabamba en Perú; 15°S marca el inicio de la Subcomisión de dominio Altiplano (Ríos - Touma et al., 2014). A modo de referencia en el mapa se ubican dos cuencas: una alta, Guayllabamba, en Ecuador; y otra media, Cañete, en Perú.

Dada la restricción de este índice para zonas andinas (> 2000 m.s.n.m.) fue preciso su aplicación considerando exclusivamente sólo los datos de las estaciones de muestreo que cumplen esta condición. Las estaciones que no se tomaron en cuenta para pruebas con el ABI en la CRP son PB3, PB2, N1 y N2 (Mapa 5).



Mapa 5. Diferencia de cotas para la CRP que se consideran como restricción en la aplicación del ABI (> 2000 m.s.n.m.).

Al existir esta gran heterogeneidad a nivel altitudinal en la CRP, una combinación de el **ABI + BMWP/Col** se llevó a cabo para ser comparada con el resto de los índices evaluados, de esta forma se consideraron todos los puntos de muestreo respetando la restricción del ABI (> 2000 m.s.n.m.). En el Mapa 5 las estaciones bajo los 2000 m.s.n.m. fueron puntos donde el BMWP/Col fue calculado mientras que en el resto el producto fue el ABI.

7.2.3.3. % EPT (Efemeróptera, Plecóptera, Trichoptera)

Es el porcentaje de la combinación de estas tres órdenes de macrozoobentos, las cuales son consideradas a nivel mundial como indicadores muy sensibles a las perturbaciones de los ríos, de tal modo la hipótesis es que siempre van a decrecer conforme existan más afecciones en los ecosistemas acuáticos. Según Barbour, et al. (1999); Crawford & Lenat, (1989), Eaton & Lenat, (1991); Quinn & Hickey, (1990), Bispo, et al. (2006), EPT es un conjunto de taxones muy sensibles prefiriendo estar siempre en aguas limpias y bien oxigenadas (Fig. 10).

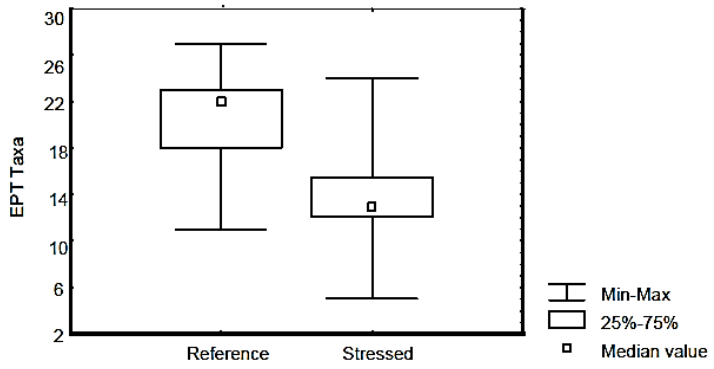


Figura 10. Potencial de discriminación del EPT en ríos sometidos a estrés ambiental (en Wyoming, E.E.U.U.; Barbour et al., 1999).

7.2.3.4. EPT / EPT + OCH (Odonatos, Coleópteros, Heteróptera)

Bonada et al. (2007, 2006) sugiere que la relación entre el número de EPT y OCH cambia a lo largo del gradiente de caudales permanentes, con una caída en la abundancia de EPT y un concomitante incremento de OCH, lo que se equipara a la variación de flujo observada desde un sistema de corrientes perennes hacia uno temporal. Un cambio en la estructura comunitaria de los macrozoobentos al inicio de la estación seca desde EPT (reofílicos mayoritariamente relacionados con áreas de rápidos) hacia OCH (representativos de sistemas lénticos) indica un claro detrimento del caudal y con ello la formación de posas que indican un avance de períodos más secos (Feminella, 1996 & Boulton, 2003). En resumen, el radio EPT/EPT + OCH refleja la temporalidad de flujo de corriente (Chaves et al., 2008), por lo tanto, es un índice adecuado para evaluar la presencia o no de estrés hídrico debido a cantidad de agua (±).

Sin embargo y aunque este índice fue concebido funcionalmente para zonas con climas mediterráneos, en Ecuador y en la CRP la diferenciación en términos de caudal es muy marcada entre la estación seca y lluviosa, sobre todo en ríos de órdenes menores. Es por ello que se ha creído conveniente hacer una aproximación inicial al uso del radio EPT / EPT + OCH y evaluar su posible utilización en ciertos ríos como los pertenecientes a la CRP.

7.2.3.5. CRP Index

Es una propuesta efectuada en este estudio que intenta combinar la tolerancia asignada (scores) con la abundancia de cada taxón en la muestra:

$$CRP\ Index = \frac{Ab_1 * Sc_1 + Ab_2 * Sc_2 + Ab_3 * Sc_3 + Ab \ldots k_{1:n} * Sc \ldots k_{1:n}}{Sc_1 + Sc_2 + Sc_3 + Sc \ldots k_{1:n}} \qquad \textbf{(1)}$$

donde, **Ab** = Abundancia del taxón; y **Sc** = Score o puntaje asignado por el índice biótico convencional a ese taxón.

Se cree que podría existir en algunos casos ciertas mejoras en términos de la optimización matemática (ajuste en un modelo de clasificación) si la variable dependiente, respuesta biológica = comunidad de macrozoobentos, no solo está dada por la sumatoria de los puntajes asignados de cada taxón, sino que también considere la abundancia de cada grupo (familia) colectada en la muestra. Por ello se propuso el cálculo de un índice a través de medias ponderadas (Ec. 1) la cual es un estadístico (medida de tendencia central) apropiado cuando en un conjunto de datos cada uno de ellos tiene una importancia relativa (peso = scores, en el presente caso) respecto de los demás datos. Se obtiene multiplicando cada uno de los datos por su ponderación (peso) para luego sumarlos, obteniendo así una suma ponderada; después se divide ésta entre la suma de los pesos, dando como resultado la media ponderada (Ec. 1).

Dado que esta propuesta matemática implica el uso de un Score (Sc) o puntaje asignado a la contaminación, primero se debió escoger entre cuál de las asignaciones disponibles funciona mejor, (primera pregunta científica) si la del ABI + BMWP/Col o el BMWP/Col, hecho esto, la asignación que mejor reconocimiento de patrones tuvo, fue la empleada para construir el CRP Index.

7.2.3.6. Muestreo de macroinvertebrados bentónicos

Los macrozoobentos fueron colectados usando una red de patada ´kick - sampling´ de 25 x 25 cm y un ojo de malla de 0,5 mm (Jacobsen et al., 1997). La muestra fue obtenida seleccionando un transecto en el lecho de río de aproximadamente 1,5 metros; este se ubicó sobre todo en los hábitats dominantes. Asimismo, con el fin de homogenizar el muestreo, se utilizó la red durante un período de 2 minutos en cada réplica. Las colectas fueron preservadas en alcohol al 85% para su posterior identificación en laboratorio hasta el taxón de familia (género en muchos casos pero para los propósitos de este trabajo la resolución taxonómica utilizada fue a nivel de familia).

Como proceso complementario al uso de la red de patada y con el fin de aumentar la eficiencia de los muestreos en términos de riqueza de taxones, con la ayuda de pinzas entomológicas por punto de colecta se revisaron de forma aleatoria piedras del lecho, colectándose así todos los organismos presentes en ellas. Este proceso finalizó al término de 20 minutos y se llevó a cabo con el fin de obtener una mayor normalización del muestreo, además de representar de forma menos sesgada la riqueza taxonómica de los sitios considerados (Roldan, 1996).

Las claves taxonómicas consideradas en este estudio fueron las de Roldan, 1996; Domínguez et al., 2006; Art & Spinelli, 2007; Stark, 2007; Coscarón & Coscarón, 2007; Heckman, 2008 & 2011 y Domínguez, 2009, entre otras.

7.2.4. Variables geomorfológicas e hidrológicas

7.2.4.1. Pendiente y altura sobre nivel del mar

Este tipo de variables permite describir algunos de los aspectos más relevantes y estables de la configuración física de las redes de drenajes y caracterizar los causes desde una perspectiva amplia (Rodríguez et al., 1997). De tal modo, la pendiente (%) y los metros sobre nivel del mar (m.s.n.m.) fueron generadas para ser utilizadas con todo el compendio global de datos, ya que ambas variables son fuertes descriptoras de la variabilidad de los macrozoobentos y de la calidad del agua en general (Jacobsen, 2004; Fig. 11).

Para obtener dicha información se utilizó la Ortofotografía del Proyecto SIG Tierras (http://servicios.sigtierras.gob.ec/) del Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP) de Ecuador. Cada ortofoto tiene una resolución de 30 x 30 cm^2 y el Modelo Digital del Terreno (MDT) de cada una de estas de 3 x 3 m^2. La pendiente se calculó en porcentaje (%).

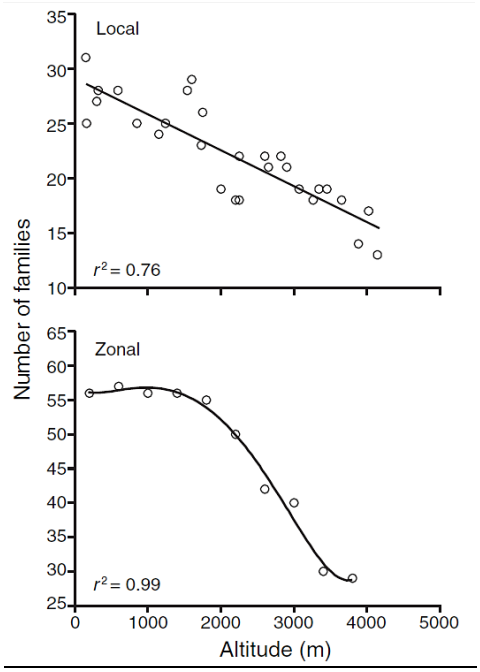


Figura 11. Regresión lineal para la riqueza local (arriba) y regresión polinómica para la riqueza zonal (abajo) de los grupos de macrozoobentos (en su mayoría familias) en relación con la altitud (m.s.n.m.) en Ecuador (Jacobsen, 2004).

7.2.4.2. Orden de río

Una importante carencia en la información levantada en la CRP para calidad de agua radica en la no medición de caudales durante los monitoreos efectuados, esto en términos de optimización de los datos conlleva principalmente a que no se logró determinar una tasa de carga contaminante peor aún efectuar procesos de modelización matemática. Todo esto ya que el caudal es un importante factor que determina la tasa de dilución de contaminantes y su marcada variabilidad estacional en zonas tropicales lo hacen muy necesario cuando se evalúan los sistemas de agua superficiales y su calidad. Con todos estos antecedentes es evidente que los sistemas biológicos acuáticos lóticos van a ser fuertemente influenciados por la cantidad de agua que fluye en el lecho, determinando el tipo de organismos y sus adaptaciones morfológicas y etológicas; por ejemplo, eventos de alto flujo pueden tener un efecto importante sobre la biomasa de algas en los arroyos tropicales y subtropicales (Pringle et al, 1986; Townsend & Padovan, 2005), así como también estos picos de crecida a menudo se asocian con un aumento de la turbidez, y con la limitación en la corriente de producción primaria (Lewis et al., 1995). Específicamente para los macrozoobentos, las densidades de estos tenderán a su punto máximo durante la estación seca, cuando los flujos son estables y se agotan por crecidas durante la estación húmeda (Dudgeon, 1996, 1998). Las diferencias de densidad entre las estaciones secas y húmedas pueden ser grandes: en el Neotrópico, las fluctuaciones de 250 a 1250 individuos por m^2 se han reportado en Río Sábalo, Costa Rica (Ramírez & Pringle, 1998b), 785 - 1672 individuos por m^2 en Río Orituco, Venezuela (Dudgeon, 2011), y de 0 a > 64500 individuos por m^2 en Río Las Marías, Venezuela (Flecker & Feifarek, 1994). Un patrón bimodal también se ha informado en pequeños arroyos de Venezuela donde los macroinvertebrados bentónicos casi se desvanecen a medida que avanza la estación seca, pero las poblaciones tienden a restablecerse rápidamente al comienzo de la temporada de lluvias (Flecker & Feifarek, 1994; Rincón & Cressa, 2000).

Así, para las preguntas de investigación planteadas los patrones estacionales no son objeto de análisis, pero sí los espaciales, tal es así que para estructurar nuestra matriz de datos y que

esta se componga de descriptores adecuados que expliquen la variabilidad de los macrozoobentos, como medida compensatoria para la falta de caudales se calculó el orden del río para cada estación de muestreo, con lo cual se pretende suplir, en cierta medida, a los caudales ya que de forma directa estos se relacionan proporcionalmente con el orden de río, el cual no es más que el grado de ramificación o bifurcación dentro de una red de drenaje (Fattorelli & Fernández, 2011). El orden del río puede determinarse de acuerdo a criterios expuestos por diferentes autores, pero en sí, todos básicamente deben reflejar cambios en las características físicas que controlan los procesos de flujo naturales según como la red se desarrolla (Andah et al. 1987). De las diferentes opciones, en este estudio se calculó el orden mediante el método de Shreve (Ferro, 2002) con la ayuda de un SIG (ArcGis versión 10.1) a través de un MDT agregando un tamaño de celda de pixel de 100 m con lo cual se buscó representar de forma adecuada las depresiones que corresponden a ríos, quebradas y arroyos en la CRP. Obtenidos los órdenes para toda la red de drenaje de cuenca, se superpuso estos para cada uno de los 64 puntos de monitoreo de calidad de agua y de forma indirecta se pudo asociar un número que implique caudal a los mismos. Al existir puntos de muestreo muy disímiles entre sí en términos de tamaño de causes y por ende de cantidad de agua, se escogió el método de Shreve ya que discretiza marcadamente las estaciones de muestro consideradas en el estudio, pues tiene un carácter acumulativo mucho mayor si se lo coteja por ejemplo con Horton - Strahler, el cual se utiliza de una manera más cotidiana. La magnitud de cualquier segmento de corriente iguala el número de la magnitud de sus fuentes, lo cual significa que Shreve es una de las relaciones más simples para predecir el flujo de corriente (Fattorelli & Fernández, 2011). El método de Shreve considera todos los enlaces de la red, así, al igual que en el método Horton - Strahler, a todos los enlaces exteriores se les asigna un orden de 1, sin embargo, para todos los enlaces interiores en el método de Shreve los órdenes son aditivos. Por ejemplo, la intersección de dos enlaces de primer orden crea un enlace de segundo orden; la intersección de un primer orden y el enlace de segundo orden crean un enlace de tercer orden y la intersección de un segundo orden y otro enlace de tercer orden crea un vínculo de quinto orden. Debido a que las órdenes son aditivos, los números del método de Shreve se refieren a veces como magnitudes en vez de órdenes, de tal forma, la magnitud de un enlace en el método de Shreve es el número de enlaces ascendentes (Tarboton et al., 1991; Fig. 12a, Fig. 12b).

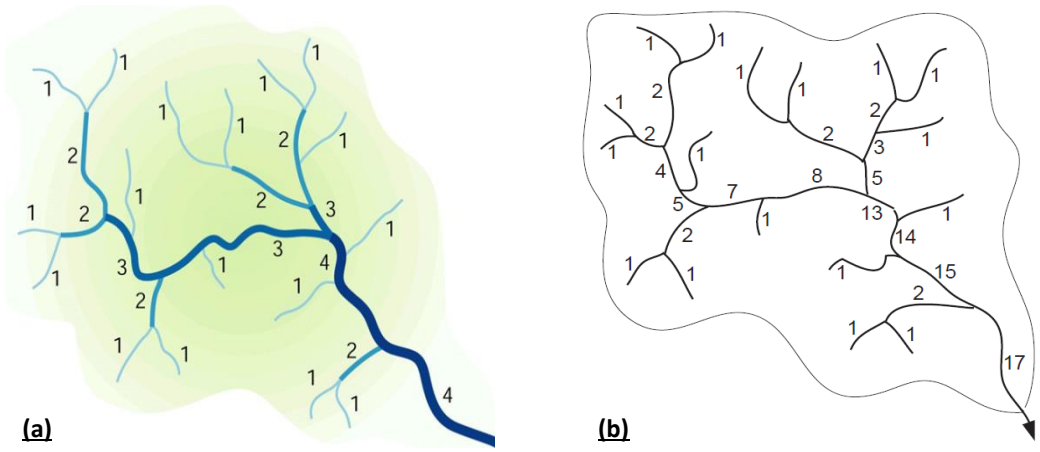


Figura 12. (a) Red de drenaje según Horton - Strahler (http://www.waterontheweb.org/) y **(b)** red de drenaje según Shreve (Bain & Stevenson, 1999).

7.3. Pre - tratamiento de datos

7.3.1. Organización de la matriz de datos

Al tratarse de datos con una alta dimensionalidad, ósea, ubicados en un espacio multivariante, se organizó un sistema matricial X en donde las filas representan a los n objetos (estaciones de muestreo para calidad de agua y sus réplicas), las cuales son descritas por p variables (columnas) del tipo físico – químicas, microbiológicas, geomorfológicas y biológicas (Tauler et al., 2009). A saber, la matriz se compone de 10234 campos ($n = 301 * p = 34$).

$$\mathbf{X} \Longrightarrow \begin{array}{c|cccc} & 1 & & & p \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \\ & x_{21} & x_{22} & \dots & x_{2p} \\ & \dots & \dots & \dots & \dots \\ n & x_{n1} & x_{n2} & \dots & x_{np} \end{array}$$

Figura 13. Esquema de la matriz ($X = n * p$) que contiene los datos de calidad de agua.

7.3.2. Valores perdidos

Los datos faltantes o valores perdidos son un problema común al momento de analizar las series de tiempo de datos tanto de calidad de agua como de hidrología, (Hirsch et al. 1982; Fattorelli & Fernández, 2011). Así, para nuestro conjunto de datos la variable oxígeno disuelto (OD) fue la que más espacios vacíos presentó a lo largo de su distribución en los muestreos realizados, con un 44,8 % de valores faltantes, lo cual se supone fue debido a errores de diseño y planificación en las campañas de muestreo. De tal forma, en casos así, la eliminación de dichas muestras no es una alternativa pues son demasiadas, y excluir la variable resulta de igual forma algo no factible al ser el OD clave en un estudio de calidad de agua, por tal motivo, se calculó los valores perdidos con un modelo de regresiones múltiple (Fig. 14) (Frank & Todeschini, 1994; Arteaga & Ferrer, 2002), en el cual la variable a computar fue explicada por varios descriptores independientes.

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots\dots \beta_n X_n + \varepsilon$$

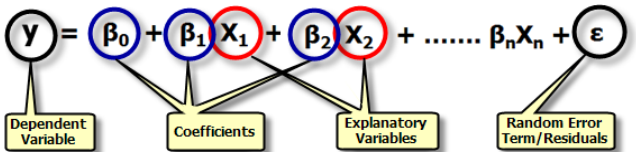


Figura 14. Esquema del análisis de regresiones múltiples (http://webhelp.esri.com/).

Los resultados muestran el ajuste de un modelo de regresión lineal múltiple para describir la relación entre OD y 8 variables independientes. La ecuación del modelo ajustado es:

$$O_2 = 9{,}21616 - 0{,}0635254 + \text{DBO} - 0{,}0117121 * \text{DQO} - 0{,}00833821 * \text{EC} - 0{,}0330213 * \text{IHF\_EPA} + 0{,}388016 * \text{pH} - 0{,}0200131 * \text{T(°C)} - 0{,}000245873 * \text{Col} - \text{F} - 0{,}000114129 * \text{Col} - \text{T}$$

Así, dado que el p - valor en el ANOVA es inferior a α, (p – valor = 0,0009 < α = 0,001) existe una relación estadísticamente significativa entre las variables para un nivel de confianza del 99%. El estadístico R^2 indica que el modelo explica un 44,07% de la variabilidad en OD, y el R^2 - ajustado que es más conveniente para comparar modelos con diferente número de variables independientes, es de 31,297%. El error estándar de la estimación muestra la desviación típica de los residuos que es 1,047. El error absoluto medio (MAE) de 0,70 es el valor medio de los

residuos. El estadístico Durbin-Watson (DW) examina los residuos para determinar si hay alguna correlación significativa basada en el orden en el que se han introducido los datos y dado que el *p* - valor es superior a 0,05, no hay indicio de autocorrelación serial en los residuos. Cabe aclarar que este modelo es ya producto de una simplificación de uno anterior, para lo cual se tuvo en cuenta los *p* - valores más altos en las variables independientes, estos términos no son estadísticamente significativos para un nivel de confianza del 90% o superior, por tanto, fueron excluidos del modelo para dejarlo tal cual como se detalla. Al analizar los resultados de cálculo para los datos faltantes, se observa que la serie obtenida está estadísticamente acorde con la distribución de OD original, aunque, existe una constante tendencia a la disminución de la dispersión de los datos obtenidos por el modelo de regresiones múltiples. Esto se debió tomar muy en cuenta en los resultados y conclusiones finales de este estudio (Fig. 15 y Fig. 16). Datos de OD obtenidos por regresiones múltiples son estadísticamente diferentes de los medidos en campo (U de Mann-Whitney; *p* - valor = 0,001 < α = 0,05). Otras variables como la DQO, DBO y la Alcalinidad, también presentaron datos faltantes en sus series de datos, empero estos casos fueron de cuatro valores por serie de datos como máximo, de modo que medidas de tendencia central como promedios y medianas se calcularon para rellenar los vacíos existentes.

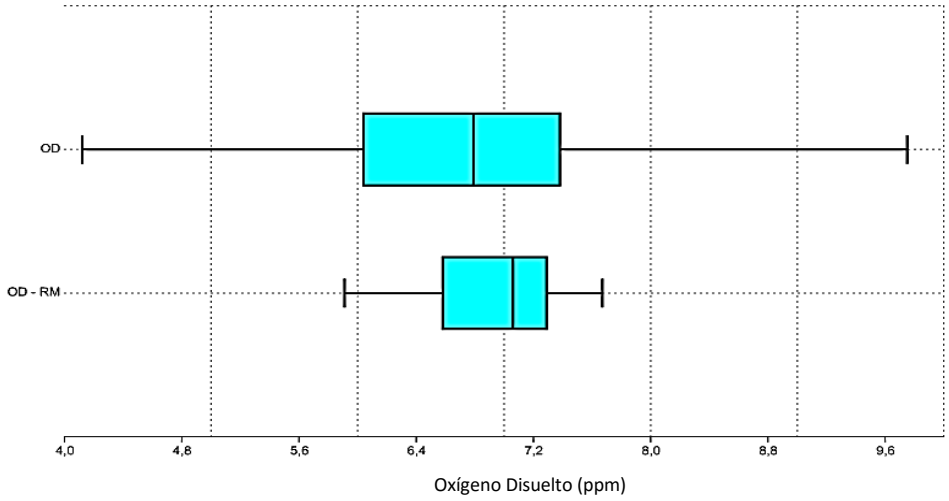


Figura 15. Boxplot de datos obtenidos con regresiones múltiples (OD - MR) y los medidos en campo (OD).

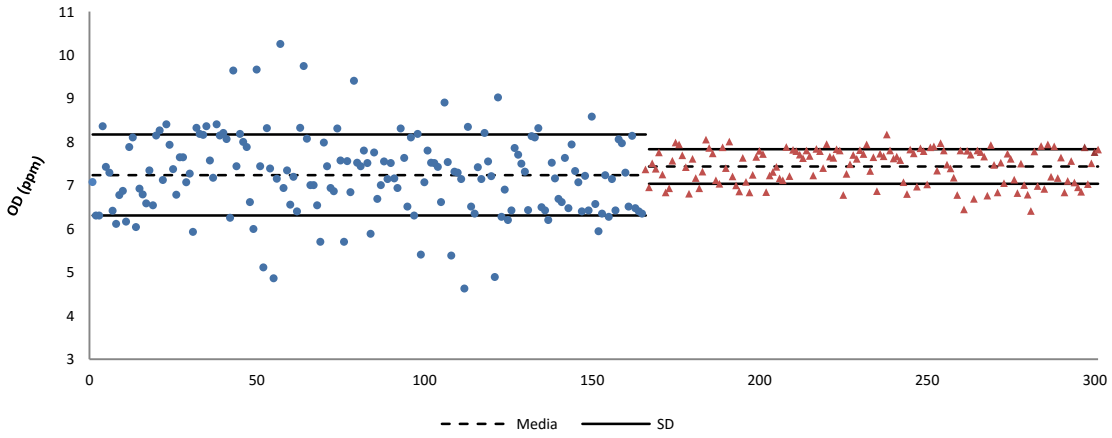


Figura 16. Círculos azules = distribución de los datos de OD medidos in situ. Triángulos rojos = datos de OD obtenidos a través del modelo de MR.

7.3.3. Unidad de análisis estadístico

Para el *k* – NN a través de GAs la unidad de análisis experimental son los objetos (estaciones de muestreo y sus réplicas) correctos y mal asignados a las clases establecidas *a priori* y condicionadas por los macrozoobentos y sus diferentes índices bióticos. Mientras más asignaciones correctas existan mejor será el ajuste en el modelo de clasificación. Por el contrario, en el PCA los objetos de análisis no son las estaciones de muestreo sino grupos de estas que se forman condicionados por el índice biótico elegido como el óptimo para la CRP, empero dichos grupos solo se conforman por variables descriptoras de tipo físico – químicas, microbiológicas y geomorfológicas, así lo que se pretende es identificar que variables del tipo descriptoras explican mayoritariamente a una u otra clase biótica.

Sin embargo, en pruebas previas se demostró que clases estándar (5 clases de calidad biótica) como las del ABI o BMWP/Col (Tabla 3), no brindaban resultados recomendables tanto en el *k* – NN + GAs como en el PCA; tal es así que el NER en el *k* – NN era bajo (μ = 44 %), y en el PCA el % de varianza explicada se distribuía en varios componentes (muchos), con lo cual la calidad de la proyección no era tan confiable (Zwanziger et al., 1997). Con estos antecedentes, y con el propósito de que los parámetros que miden el rendimiento de nuestros métodos estadísticos multivariantes utilizados alcancen niveles satisfactorios, arbitrariamente se llevó a cabo pruebas que consistieron en reformular los bordes o fronteras de las clases bióticas. Finalmente, a las series de datos de los índices de macrozoobentos, se las dividió en sus percentiles[6] 33,33 % y 66,66 % obteniéndose como tal solo tres clases de calidad biótica (Tabla 4, Fig. 17). Este proceso de re categorización de las clases bióticas también obedeció (aunque en una mínima parte) a que medidas como nuestra propuesta CRP Index no posee una división de clases como el ABI o el BMWP/Col.

Tabla 4. Nueva clasificación de los índices de macrozoobentos.

Percentil (%)	Clase	Significado	Color
≥ 66.66	I	Aguas muy limpias a limpias, o muy pocos efectos de contaminación	
≥ 33.33; 66.66 ≤	II	Evidentes algunos efectos de contaminación	
≤ 33.33	III	Aguas contaminadas a muy contaminadas	

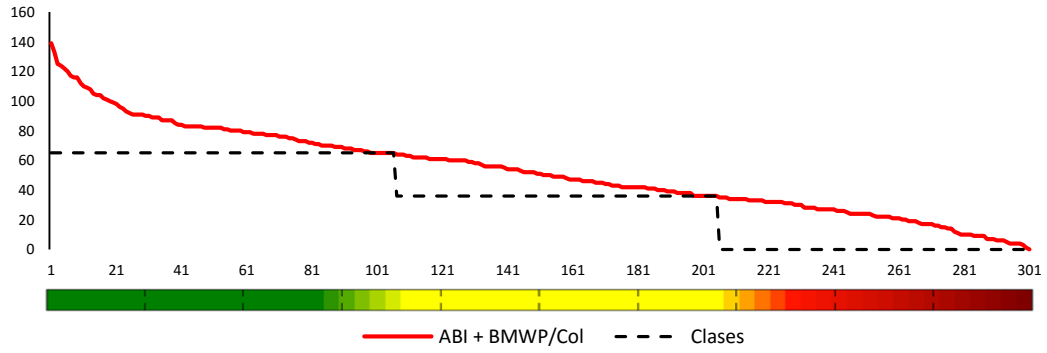


Figura 17. Ejemplo de la nueva redistribución de clases bióticas dadas por el cálculo de percentiles para la serie de datos del ABI + BMWP/Col.

[6] Es un estadístico de posición el cual va a tomar valores de la variable caracterizados por superar a cierto porcentaje de observaciones en la población (o muestra). Para una variable discreta, se define el percentil de orden *k*, como la observación, P*k*, que deja por debajo de si el *k* % de la población (Díaz et al., 2010).

El planteamiento y utilización de estas tres clases mejoraron significativamente los parámetros de evaluación tanto del *k* – NN + GAs como del PCA y se constituyó en un interesante aporte para la CRP, manifestando así que a nivel regional aún existen muchas necesidades de investigación que deben ser atendidas con el propósito de lograr mejoras sustanciales en el manejo y gestión de los recursos hídricos superficiales. Asimismo, queda en evidencia que al ser la CRP un sistema hídrico muy heterogéneo y con características muy intrínsecas, herramientas de gestión y elementos de juicio que han sido elaborados en otras latitudes deben de ser manejados y aplicados con sumo cuidado.

7.3.4. Análisis de datos (marcha metodológica)

En un orden cronológico se detalla a continuación la marcha metodológica empleada para nuestro tratamiento y análisis de datos:

a. Se calcularon los datos faltantes de OD a través de un modelo de regresiones múltiples y se organizó la matriz de datos ($X = n * p$).
b. Se computaron los diferentes índices bióticos a través de la comunidad de macrozoobentos y a los mismos se les calcularon sus percentiles 33,33 % y 66,66 % con el fin de que todos brinden tres clases de calidad de agua (Tabla 4, Fig. 17).
c. Modelos de clasificación a través del algoritmo *k* – NN + GAs fueron construidos, en donde a cada uno le correspondió una variable de respuesta biológica distinta que incumbe a cada uno de los índices bióticos comparados (Fig. 18). Los resultados en la asignación de patrones con menor ER fueron los que permitieron escoger el índice biótico más adecuado.

    - Las fronteras o bordes de las clases de los objetos analizados (por ende también de los índices bióticos estudiados) se corrigieron en un proceso conocido como 'editing' (Wilson, 1972)[7], esto con el fin de mejorar el modelo de clasificación (menor ER , mayor precisión y más alto poder en predicción), de tal modo al momento de escoger cual sería el mejor índice biótico no solo se utilizó el criterio del menor ER sino también (muy importante) cual prueba necesitó menos 'editing', pues cuanto mayor sea este el modelo resultante presentará un ajuste solo para un determinado conjunto de datos homogenizados (con el consecuente bajo poder predictivo del modelo). Pruebas pre clasificatorias se llevaron a cabo y se identificaron objetos 'outliers' causantes de ruido y fueron suprimidos. Arbitrariamente se decidió que por cada prueba realizada para cada índice biótico evaluado un máximo de un 15 % de objetos 'outliers' podrían eliminarse a través del proceso de 'editing' (retoque de datos).

[7] Es la eliminación de objetos o estaciones de muestreo que en las pruebas iniciales del *k* – NN con GAs eran mal asignadas, es decir, la eliminación de datos atípicos 'outliers' que tienen efectos de ruido en el proceso de modelización. A mayor proceso de 'editing' el poder predictivo del modelo decrece pues este se constituye en un instrumento matemático puntual para un conjunto específico de datos forzadamente homogenizados, siendo esto en un sentido de aplicabilidad, algo irreal (Zhi & Zhou, 2004).

- Muchos enfoques han sido propuestos con el fin de mejorar los resultados en la correcta asignación de patrones en un modelo de clasificación (Jiang & Zhou, 2004). El primero de estos planteado por Wilson (1972) es el que se llevó a cabo en el presente caso y básicamente en esta regla el 'editing' de la serie de referencia se realiza en primer lugar y luego, todas las muestras en el conjunto de referencia se clasifican utilizando la regla *k* - NN; las muestras clasificadas erróneamente se eliminan del conjunto de referencia para que luego cualquier muestra de entrada sea clasificada utilizando la regla *k* - NN y el conjunto de referencia editado (Chang et al., 2011; Fig. 21).

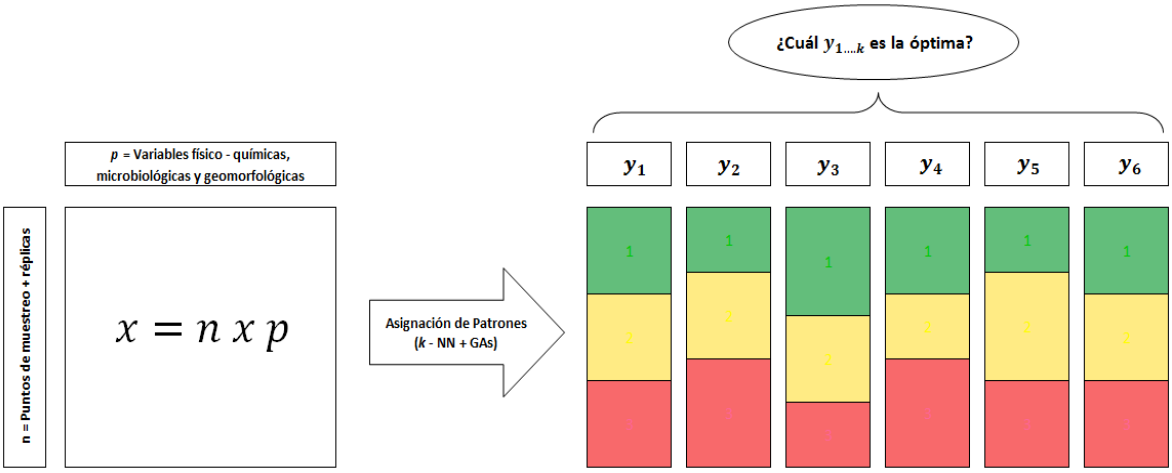


Figura 18. Esquema de la primera pregunta planteada. ¿Cuál de las variables de repuesta biológica dada por los macrozoobentos ($\boldsymbol{y_{1,2,3,4,5,6}}$) es la óptima en términos de un ajuste matemático para un modelo de clasificación? Cada $\boldsymbol{y}$ corresponde a los distintos índices bióticos evaluados; el mejor de estos fue al que le correspondieron la mayoría de objetos (descritos por $\boldsymbol{X}$) correctamente asignados.

d. Una vez seleccionada la medida de macrozoobentos óptima y funcional para la CRP, a los datos de las variables descriptoras ($\boldsymbol{X}$) se los subdividió en las tres clases que dictamina el índice escogido, de tal forma, un PCA se llevó a cabo entre estos tres grupos de datos ($\boldsymbol{X}$). Por obvias razones la variable de respuesta biológica no se incluyó en el PCA de forma directa, sino solo al diferenciar los grupos por los colores representativos de cada clase (Fig. 19, Fig. 20).

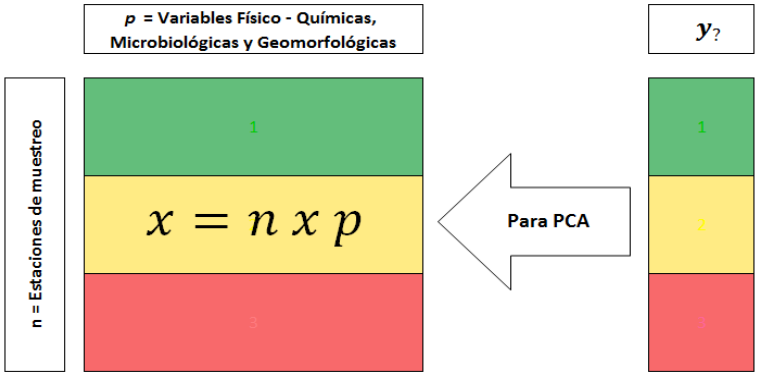


Figura 19. Esquema de la segunda pregunta planteada. $y_?$ = índice biótico escogido como óptimo, sus bordes de clases determinan los nuevos tres grupos de las variables descriptoras que serán sometidas al PCA, lo que permite explorar cuales son las variables que explican mayoritariamente la variabilidad biótica de la CRP.

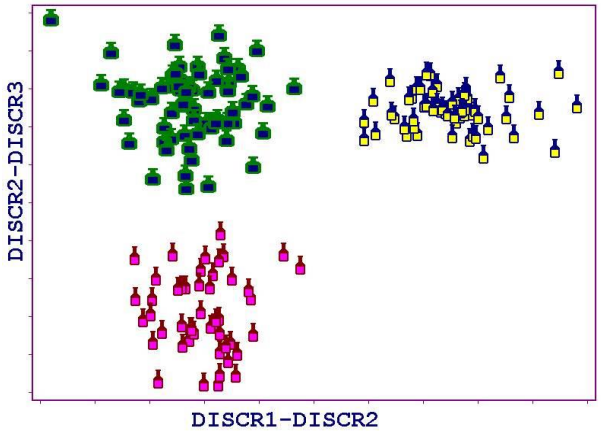


Figura 20. Referente de los grupos de análisis bajo los cuales se llevó a cabo el PCA, se detalla un ejemplo análogo al presente caso. (Apuntes de clases del Dr. Roberto Todeschini del Milano Chemometrics & QSAR Research Group.

En la Fig. 20 se detallan tres grupos de diferentes tipos de vinos y para el presente caso cada botella equivaldría a una estación de muestreo y cada grupo a una clase de calidad de agua dictaminada por los macrozoobentos.

e. En el PCA, solo los pesos iguales o mayores a 0,2 fueron los elegidos (Bere & Tundisi, 2011) para determinar cuáles parámetros de los originales describen apropiadamente la variabilidad de las tres clases (índice biótico).

f. **Validación del PCA.** La fiabilidad científica del PCA debe ser validada a través del uso de otros métodos independientes, y una forma de lograr este objetivo es comparar los datos de calidad del agua con y sin las variables que el PCA determina como 'no principales' ($Pesos < 0{,}2$). De tal modo, análisis de regresiones múltiples (MR) se llevaron a cabo, uno inicial con el índice biótico determinado como óptimo 'variable dependiente' respecto de todas las 33 variables descriptoras 'variables independientes' y otro, con la misma variable dependiente empero respecto solo de las variables seleccionadas como principales por el PCA (Ouyang, 2005). Teóricamente, si las variables decretadas como 'no principales' ($Pesos < 0{,}2$) por el PCA son excluidas del análisis de MR, el R^2 - ajustado[8] no debería de variar significativamente entre una prueba y otra.

7.3.5. Paquetes informáticos utilizados

Para el *k* – NN + GAs el software utilizado fue el entorno de programación **MATLAB**, y específicamente las extensiones **Classification toolbox 3.1** (para el 'editing') y **ga toolbox** (para el *k* – NN + GAs), las cuales son una colección de módulos para obtener modelos multivariados en clasificación, que han sido desarrollados por el Milano Chemometrics and QSAR research Group del Dipartimento di Scienze dell'Ambiente e del Territorio e di Scienze della Terra, Università degli Studi di Milano - Bicocca (Ballabio & Consonni, 2013).

Referente al Análisis de Componentes Principales (PCA) el software utilizado fue el **PAST: PALEONTOLOGICAL STATISTICS SOFTWARE PACKAGE 2.17** el cual es un software libre para el análisis de datos científicos, con funciones de manipulación de datos, gráficos, estadísticas univariantes y multivariantes, análisis ecológico, series de tiempo y análisis espacial, morfometría y para estratigrafía. Ha sido desarrollado por el Natural History Museum, de la University of Oslo (Hammer et al., 2001). Con este software el PCA se puede llevar a cabo a nivel de grupos de datos predeterminados en gabinete y condicionados, para nuestro caso, por el índice biótico elegido como óptimo. El análisis se desarrolló entre las medias de los grupos (es decir, los elementos analizados fueron los grupos, no las filas o puntos de muestreo para nuestro caso) y los puntajes de PCA se calcularon utilizando el vector de productos con los datos originales (Reisenhofer et al. 1998; Hammer, Harper & Ryan, 2007).

Para construir los modelos de regresiones múltiples empleados en la validación del PCA y para el cálculo de valores faltantes de OD, el paquete informático utilizado fue una versión Demo del paquete estadístico **STATGRAPHICS 5.1** el cual es un potente programa intuitivo para el análisis y visualización de datos, modelización y de análisis predictivos.

[8] El R^2 - ajustado indica el poder explicativo de los modelos de regresión que contienen diferente número de predictores. Es decir, es más conveniente para comparar modelos con diferente número de variables independientes (Rawlings et al., 1998).

8. RESULTADOS

8.1. Para la primera pregunta científica planteada

Los GAs han sido usados a nivel mundial en nexo con otras técnicas de reconocimiento de patrones y para este caso su uso estuvo enfocado en la optimización de los modelos de clasificación obtenidos a través del algoritmo *k* – NN (He et al., 1999; Suguna & Thanushkodi, 2010). La marcha metodológica a través del *k* – NN + GAs y el script y las opciones empleadas en el MATLAB para su extensión 'ga toolbox' se detallan a continuación:

```
>> data=[Las estaciones de muestreo y sus réplicas junto con los datos de las
variables que las describen (geomorfológicas, físico – químicas y microbiológicas)];

>> y=[Las tres clases de calidad de agua condicionadas por los percentiles 33.33 % y 66.66 %
(esta respuesta biológica es la que cambia según los distintos índices bióticos y sus variantes)];

prep ------------------------------------------------> % Preprocesamiento de datos.
no=0;cetering=1;autoscaling=22                         % Crea un conjunto de entrenamiento de los
X sin transf.:0;X=X2:10                                archivos particionados de la "y" biunívoca.
>> options = ga_options('knn','none',100)

options =

    method: 'knn'
      scal: 'autoscaling'
 cv_groups: 5
   cv_type: 'vene' -------------------------------------> Tipo de crossvalidation
 num_chrom: 30 --------------------------------------> Número de cromosomas
  startvar: 5 ---------------------------------------> Número de variables por cromosoma en la población inicial
    maxvar: 30 --------------------------------------> Número máximo de variables en cada cromosoma
   probmut: 0.0100 -----------------------------------> Probabilidad de mutación
 probcross: 0.5000 -----------------------------------> Probabilidad de cross - over (50%)
      runs: 100 ----------------------------------------> Número de GA corridas
  num_eval: 100 ----------------------------------------> Número de evaluaciones
    kernel: 'linear'
class_prob: 2
 freq_back: 100
num_windows: 1
 dist_type: 'euclidean' ------------------------------> Tipo de distancia
     domax: 0

   >> res = ga_model(XXtrain,ytrain,options,0)
```

Así, en primer orden y como base al proceso de 'editing' (Wilson, 1972) se había señalado con anterioridad en el texto que, arbitrariamente, se decidió que sea cual fuere el resultado de ER por modelo generado, máximo se podría eliminar hasta un 15 % de los objetos. De este modo cada modelo dado por cada medida biótica no tendría un peso mayor o menor en términos de 'editing' y por tanto su comparación sería factible. Dentro de este mismo concepto es lógico suponer que dicho porcentaje máximo de objetos para los cuales cabría su eliminación también obedeció al criterio que a mayor pérdida de objetos el poder predictivo del modelo y el ajuste del mismo solo estarían dados para un conjunto de datos forzadamente homogenizados y muy puntuales, por ende, su uso sería poco adaptable a la realidad de la CRP. Este 15 % máximo permisible de 'editing' fue un discernimiento en aras de un resguardo con el fin de evitar en la medida de lo posible efectos de 'ink blot' (Tauler et al., 2009) en este trabajo.

Se demostró que en el proceso de 'editing' llevado a una segunda etapa, excede marcadamente el criterio de máximo un 15% de eliminación de 'outliers', (he inclusive en parte de la primera fase; Fig. 21); así, para el propósito del presente trabajo los modelos de clasificación sobre los cuales se elaboraron los juicios de selección / elección fueron los que corresponden con la primera etapa de eliminación de 'outliers' (barras verdes y amarillas; Fig. 21).

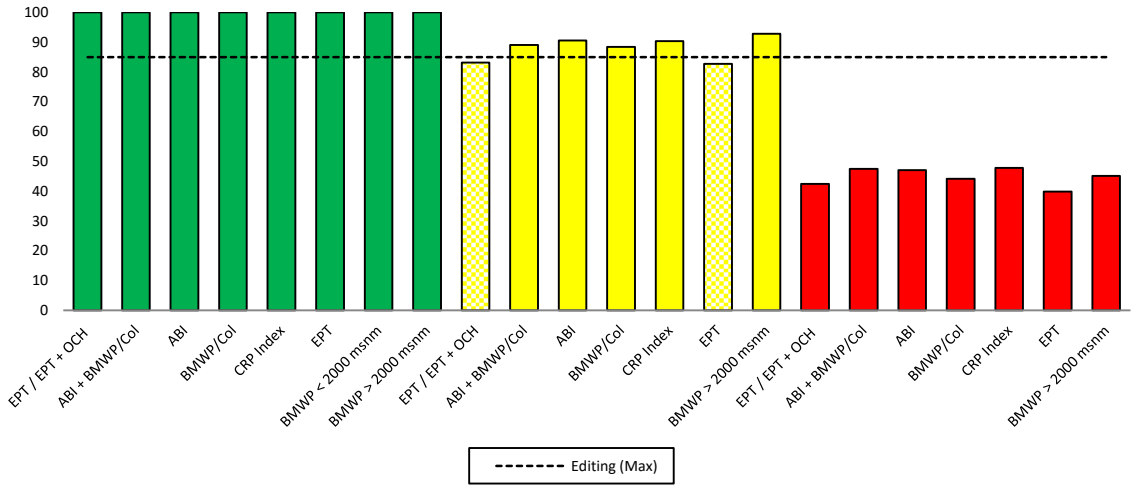


Figura 21. Porcentaje de información empleada en el análisis. En verde, para cada índice biótico evaluado y sus variantes se detallan el 100 % de los datos sometidos a un análisis de clasificación *k* – NN + GAs. En amarillo, se muestran los datos luego del primer proceso de 'editing'. En rojo, se muestran los resultados para una segunda y última etapa de 'editing'.

A saber, para cada etapa de 'editing' se describen los modelos de clasificación (*k* - NN + GAs) generados a partir de los distintos índices bióticos (y sus variantes).

Tabla 5. Modelos de clasificación calculados con las distintas medidas bióticas y sus variantes en la etapa de sin proceso de 'editing' (100 % de los datos utilizados). Una síntesis de las características principales de los modelos también se especifican (Promedios de NER y su desviación estándar (SD), las principales variables seleccionadas así como la frecuencia de estas en los modelos); FSS = Final stepwise selection.

Editing	Índice Biótico	FSS	Modelo de Clasificación (*k* - NN + GAs)																				NER (μ)	NER (SD)
			Peso de Variables																						
Con el 100 % de los datos	EPT / EPT+OCH	10	Variables Seleccionadas	msnm	Shreve	P	Alkalinity	EC	Dureza	PO_4	Al	Ca	CL-	-	-	-	-	-	-	-	-	-	-	0,52	0,03
			NER	0,44	0,49	0,48	0,49	0,53	0,52	0,53	0,56	0,56	0,54	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	53	43	39	33	32	31	28	25	24	21	-	-	-	-	-	-	-	-	-	-		
	ABI+BMWP / Col	16	Variables Seleccionadas	IHF - EPA	Cd	msnm	Col - T	NH4+	Alkalinity	Mg	Pb	Dureza	EC	Fe	CL-	T(°C)	DBO	DQO	Ca	-	-	-	-	0,59	0,04
			NER	0,48	0,5	0,61	0,61	0,62	0,63	0,63	0,66	0,65	0,63	0,61	0,59	0,59	0,58	0,59	0,59	-	-	-	-		
			Frecuencia de Selección	73	59	37	31	28	24	23	21	20	19	18	15	14	13	11	10	-	-	-	-		
	ABI	6	Variables Seleccionadas	IHF - EPA	OD	T(°C)	Dureza	CL-	F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,62	0,03
			NER	0,56	0,61	0,63	0,64	0,64	0,63	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	97	48	21	19	16	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	BMWP/Col	19	Variables Seleccionadas	IHF - EPA	T(°C)	msnm	Fe	NO3-	K	Col - F	Alkalinity	Pb	DQO	Cu	DBO	OD	TS	F-	Mg	P	EC	Cd	Dureza	0,52	0,03
			NER	0,47	0,45	0,46	0,49	0,51	0,53	0,54	0,52	0,54	0,54	0,54	0,55	0,55	0,54	0,55	0,52	0,54	0,55	0,49	-		
			Frecuencia de Selección	56	47	46	36	30	27	26	24	22	21	19	18	17	15	13	12	11	10	7	-		
	CRP Index	6	Variables Seleccionadas	msnm	Alkalinity	Ca	OD	Fe	P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,49	0,04
			NER	0,42	0,5	0,48	0,5	0,56	0,53	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	56	44	39	37	35	34	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	EPT	8	Variables Seleccionadas	IHF - EPA	msnm	DQO	DBO	Col - F	Al	K	Pb	-	-	-	-	-	-	-	-	-	-	-	-	0,56	0,03
			NER	0,48	0,56	0,55	0,56	0,59	0,58	0,60	0,57	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	91	74	32	26	24	21	20	18	-	-	-	-	-	-	-	-	-	-	-	-		
	BMWP < 2000 msnm	3	Variables Seleccionadas	IHF - EPA	DQO	Fe	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,68	0,08
			NER	0,57	0,71	0,76	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	62	40	39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	BMWP > 2000 msnm	3	Variables Seleccionadas	IHF - EPA	P	T(°C)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,57	0,03
			NER	0,53	0,58	0,6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	93	35	33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		

Tabla 6. Modelos de clasificación calculados con las distintas medidas bióticas y sus variantes en la primera etapa de 'editing'. Una síntesis de las características principales de los modelos también se especifican (Promedios de NER y su desviación estándar (SD), las principales variables seleccionadas así como la frecuencia de estas en los modelos); FSS = Final stepwise selection.

Editing	Índice Biótico	FSS	Modelo de Clasificacion (*k* - NN + GAs)											NER (μ)	NER (SD)
			Peso de Variables en los modelos												
Luego de la 1ra Etapa	EPT / EPT+OCH	9	Variables Seleccionadas	msnm	OD	IHF - EPA	EC	Alkalinity	Pb	Cu	Shreve	NO3-	-	0,65	0,03
			NER	0,55	0,62	0,66	0,67	0,64	0,64	0,65	0,69	0,69	-		
			Frecuencia de Selección	79	62	48	38	35	32	30	26	21	-		
	ABI+BMWP /Col	8	Variables Seleccionadas	Pb	IHF - EPA	NH4+	Shreve	OD	T(°C)	EC	Fe	-	-	0,6	0,06
			NER	0,54	0,48	0,63	0,61	0,67	0,63	0,63	0,65	-	-		
			Frecuencia de Selección	45	42	41	35	32	31	29	26	-	-		
	ABI	4	Variables Seleccionadas	IHF - EPA	Fe	Cu	TS	-	-	-	-	-	-	0,63	0,03
			NER	0,61	0,61	0,63	0,68	-	-	-	-	-	-		
			Frecuencia de Selección	97	27	23	22	-	-	-	-	-	-		
	BMWP/Col	8	Variables Seleccionadas	IHF - EPA	Mg	Shreve	Alkalinity	Dureza	Cu	T(°C)	msnm	-	-	0,63	0,04
			NER	0,59	0,56	0,6	0,65	0,66	0,66	0,69	0,63	-	-		
			Frecuencia de Selección	94	57	44	43	37	28	27	25	-	-		
	CRP Index	10	Variables Seleccionadas	T(°C)	OD	Fe	IHF - EPA	Alkalinity	UNT	Col - F	Mg	K	DQO	0,53	0,03
			NER	0,48	0,51	0,51	0,54	0,48	0,51	0,56	0,57	0,55	0,54		
			Frecuencia de Selección	58	43	42	38	32	31	26	25	24	23		
	EPT	5	Variables Seleccionadas	msnm	DQO	OD	DBO	UNT	-	-	-	-	-	0,54	0,06
			NER	0,43	0,53	0,56	0,58	0,61	-	-	-	-	-		
			Frecuencia de Selección	70	52	43	40	38	-	-	-	-	-		
	BMWP > 2000 msnm	3	Variables Seleccionadas	IHF - EPA	Col - T	P	-	-	-	-	-	-	-	0,58	0,02
			NER	0,55	0,58	0,60	-	-	-	-	-	-	-		
			Frecuencia de Selección	96	42	37	-	-	-	-	-	-	-		

Tabla 7. Modelos de clasificación calculados con las distintas medidas bióticas y sus variantes en la segunda etapa de 'editing'. Una síntesis de las características principales de los modelos también se especifican (Promedios de NER y su desviación estándar (SD), las principales variables seleccionadas así como la frecuencia de estas en los modelos); FSS = Final stepwise selection.

Editing	Índice Biótico	FSS	Modelo de Clasificación (k - NN + GAs)																				NER (μ)	NER (SD)
			Peso de Variables																					
Luego de la 2da Etapa	EPT / EPT+OCH	8	Variables Seleccionadas	msnm	Alkalinity	F-	IHF - EPA	CL-	Fe	Cd	NO3-	-	-	-	-	-	-	-	-	-	-	-	-	0,81	0,06
			NER	0,70	0,78	0,78	0,78	0,81	0,86	0,89	0,87	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	76	67	65	45	39	35	34	31	-	-	-	-	-	-	-	-	-	-	-	-		
	ABI+BMWP /Col	8	Variables Seleccionadas	IHF - EPA	Col - T	Alkalinity	Shreve	Fe	Mg	OD	NH4+	-	-	-	-	-	-	-	-	-	-	-	-	0,74	0,05
			NER	0,68	0,65	0,70	0,73	0,78	0,78	0,79	0,80	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	62	60	57	51	39	36	34	33	-	-	-	-	-	-	-	-	-	-	-	-		
	ABI	20	Variables Seleccionadas	T(°C)	Fe	Col - F	EC	Ca	Shreve	Mg	Alkalinity	Dureza	NH4+	CL-	P	PO_4	pH	K	Cu	T(°C)	OD	msnm	Ni	0,82	0,06
			NER	0,71	0,69	0,73	0,73	0,8	0,8	0,81	0,84	0,83	0,83	0,84	0,85	0,87	0,88	0,89	0,89	0,87	0,85	0,85	0,83		
			Frecuencia de Selección	86	60	53	38	37	36	33	32	29	26	24	23	22	19	18	17	16	15	14	11		
	BMWP/Col	8	Variables Seleccionadas	IHF - EPA	Alkalinity	Dureza	msnm	Col - T	T(°C)	T(°C)	OD	-	-	-	-	-	-	-	-	-	-	-	-	0,76	0,05
			NER	0,66	0,7	0,7125	0,78	0,78	0,8	0,83	0,81	-	-	-	-	-	-	-	-	-	-	-	-		
			Frecuencia de Selección	99	64	34	33	29	27	26	23	-	-	-	-	-	-	-	-	-	-	-	-		
	CRP Index	16	Variables Seleccionadas	msnm	OD	Col - F	T(°C)	Mg	Dureza	Alkalinity	Slope	DBO	NO3-	Col - F	Fe	Shreve	P	T(°C)	UNT	-	-	-	-	0,77	0,08
			NER	0,57	0,68	0,64	0,67	0,68	0,75	0,80	0,83	0,79	0,84	0,84	0,84	0,86	0,84	0,83	0,83	-	-	-	-		
			Frecuencia de Selección	73	69	40	39	37	34	32	30	28	27	26	24	23	22	18	17	-	-	-	-		
	EPT	14	Variables Seleccionadas	msnm	DQO	Dureza	OD	PO_4	DBO	IHF - EPA	Ca	Col - T	T(°C)	Mg	P	Shreve	Col - F	-	-	-	-	-	-	0,75	0,06
			NER	0,65	0,61	0,75	0,68	0,75	0,75	0,74	0,79	0,79	0,81	0,76	0,78	0,78	0,79	-	-	-	-	-	-		
			Frecuencia de Selección	90	55	42	38	36	32	30	28	27	26	24	23	21	17	-	-	-	-	-	-		
	BMWP > 2000 msnm	17	Variables Seleccionadas	IHF - EPA	Col - T	P	Alkalinity	PO_4	Dureza	Cu	TS	Ca	NH4+	Pb	Col - F	UNT	CL-	DBO	Mg	Slope	-	-	-	0,78	0,04
			NER	0,67	0,72	0,72	0,78	0,76	0,76	0,76	0,79	0,78	0,82	0,81	0,79	0,83	0,83	0,81	0,82	0,82	-	-	-		
			Frecuencia de Selección	85	72	46	37	33	31	30	28	27	25	20	19	18	17	15	14	13					

FSS = Final stepwise selection o 'Etapas de selección final' ya es un producto como tal del análisis del *k* – NN + GAs y determina el número de variables óptimo para la construcción del modelo de clasificación. Es decir, la cantidad de variables para cuando el mayor NER es posible. Los resultados que se logran a través del MATLAB y su extensión 'ga toolbox' dan un listado de las variables escogidas dentro del FSS así como su frecuencia de uso y sus respectivos valores de NER (Tablas 5, 6, 7).

Se hace referencia a la primera aproximación efectuada como objetivo de estudio, acerca de ¿cuál índice biótico dado por la comunidad de macrozoobentos es el óptimo como medida de respuesta biológica para un modelo de clasificación? De tal manera, se ha encontrado que la combinación del ABI + BMWP/Col resultó el parámetro biótico más eficiente para determinar las clases de calidad de agua en la CRP.

En este contexto, la combinación ABI + BMWP/Col y su nombramiento como la medida biótica óptima y operativa para la CRP resulta de varias consideraciones.

En primer lugar, dada la restricción del ABI de ser usado solo en zonas ubicadas arriba de los 2000 m.s.n.m. (Ríos-Touma et al., 2014), pudo ser aplicable en el 88,4 % de los objetos analizados (estaciones de muestreo y sus réplicas) pues el otro porcentaje (11,6 %) corresponde a puntos de monitoreo menores a los 2000 m.s.n.m.; así, el ABI como tal se constituye en un excelente instrumento biótico de evaluación [sin 'editing' el NER (μ) = 0,618], pero parcial para la CRP (cota mínima = 440 m.s.n.m.; cota máxima = 4637 m.s.n.m.) y por lo tanto necesitó ser complementado. En contexto, los otros índices evaluados (excepto el BMWP/Col) no podían contemplarse como esta parte adicional del ABI en la CRP pues aunque proceden todos de la comunidad de macrozoobentos, sus distribuciones numéricas obedecen a marchas matemáticas distintas, solamente el ABI y el BMWP/Col provienen únicamente de la asignación de scores a las familias de macroinvertebrados bentónicos (puntajes del 1 al 10) por ende pudieron ser equiparados. Sin embargo, el ABI + BMWP/Col y su elección como medida más adecuada no solo obedeció a esta compatibilidad matemática sino también a que los modelos construidos a partir de ambos índices, principalmente en su forma individual, fueron los que conllevaron una menor cantidad de eliminación de 'outliers' (Fig. 21) y a su vez un mayor NER (*k* – NN + GAs) (Tablas 5 y 6; Fig. 22).

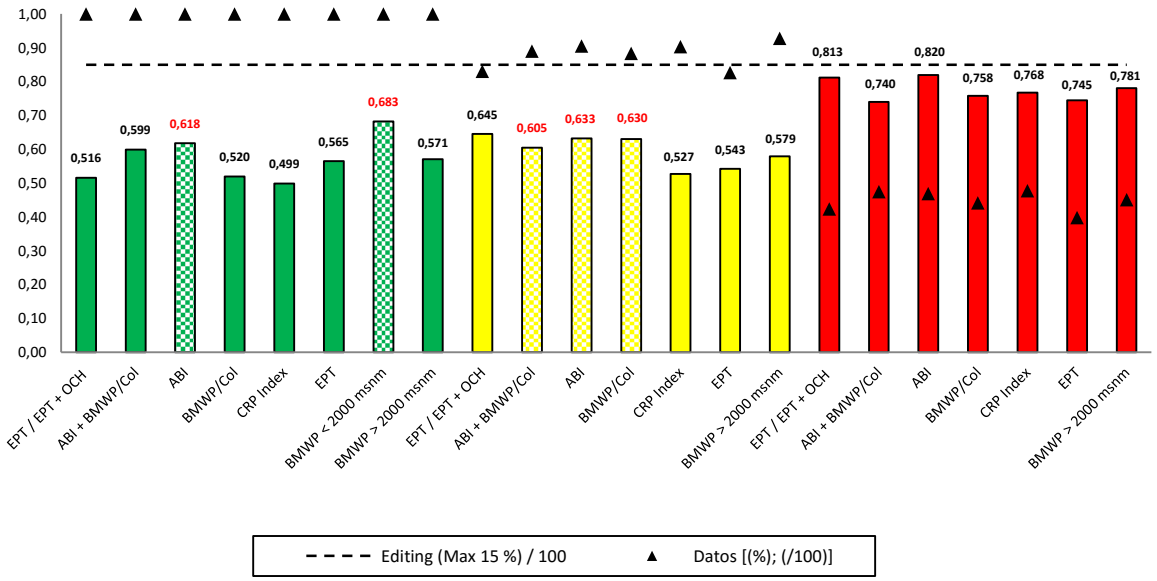


Figura 22. Verde, amarillo y rojo y los triángulos negros representan las etapas de 'editing' que se efectuaron. El NER (µ) obtenido de cada modelo son los valores ubicados arriba de cada barra. Para efectos de esta figura los porcentajes de objetos (estaciones de muestreo y sus réplicas) que fueron calculados a través del 'editing' y utilizados para la construcción del modelo (*k* – NN + GAs) se presentan como fracción en una misma escala que el resto de variables (triángulos negros).

En la Fig. 22 existen datos de precisión [NER (µ)] que están marcados con rojo, estos corresponden a los mejores modelos. Los criterios para ser señalados como tales básicamente obedecen a que resultan de un proceso de 'editing' moderado o nulo y a valores de NER (µ) mayores al 60 %.

En síntesis, los criterios de precisión (NER) y un 'editing' sensato [(no forzado); (Máximo 15 %)] tuvieron que equilibrarse en el plano de decisiones y de los múltiples modelos que se generaron el que tiene las clases (*a priori*) bióticas óptimas / operativas para la CRP es el producto de la combinación ABI + BMWP/Col.

Tabla 8. Resumen de las Tablas 5 y 6 que detalla los mejores modelos de clasificación obtenidos.

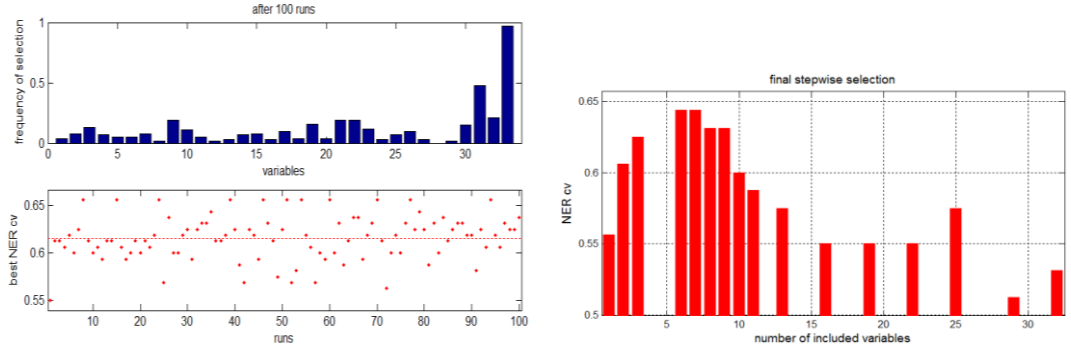
Editing	Índice Biótico	FSS	Modelo de Clasificación (k - NN + GAs)								NER (µ)	NER (SD)
			Peso de Variables									
Con el 100 % de los datos	ABI	6	Variables Seleccionadas	IHF - EPA	OD	T(°C)	Dureza	CL-	F-	-	-	0,62	0,03
			NER	0,56	0,61	0,63	0,64	0,64	0,62	-	-		
			Frecuencia de Selección	97	48	21	19	16	15	-	-		
	BMWP < 2000 msnm	3	Variables Seleccionadas	IHF - EPA	DQO	Fe	-	-	-	-	-	0,68	0,08
			NER	0,57	0,71	0,8	-	-	-	-	-		
			Frecuencia de Selección	62	40	39	-	-	-	-	-		
Luego de la 1ra Etapa	ABI+BMWP /Col	8	Variables Seleccionadas	Pb	IHF - EPA	NH4+	Shreve	OD	T(°C)	EC	Fe	0,6	0,04
			NER	0,54	0,5	0,63	0,61	0,7	0,63	0,63	0,65		
			Frecuencia de Selección	45	42	41	35	32	31	29	26		
	ABI	4	Variables Seleccionadas	IHF - EPA	Fe	Cu	TS	-	-	-	-	0,63	0,02
			NER	0,61	0,61	0,63	0,7	-	-	-	-		
			Frecuencia de Selección	97	27	23	22	-	-	-	-		
	BMWP/Col	8	Variables Seleccionadas	IHF - EPA	Mg	Shreve	Alkalinity	Dureza	Cu	T(°C)	msnm	0,63	0,04
			NER	0,59	0,56	0,6	0,7	0,7	0,7	0,7	0,63		
			Frecuencia de Selección	94	57	44	43	37	28	27	25		

Dentro del contexto global de este trabajo, el *k* – NN a través de GAs también brinda resultados sobre cuáles variables y su frecuencia de uso utiliza el modelo para obtener un ajuste máximo posible (Tablas 5, 6, 7 y 8), empero resultados de esta naturaleza se proyectaron responder desde un inicio mediante el uso del PCA (segunda pregunta), de tal

modo en esa sección del presente trabajo se analizará también los datos suministrados por el *k* – NN + GAs referente a las variables descriptoras elegidas como principales.

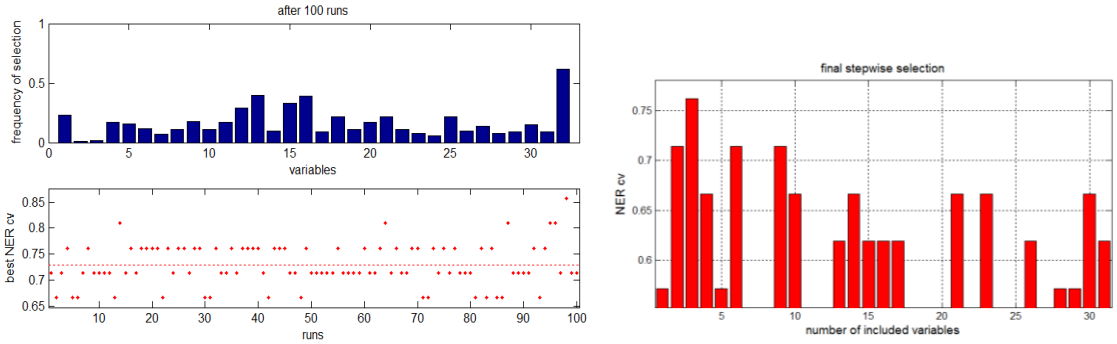
En la Tabla 8 están en compendio los mejores modelos de clasificación obtenidos a través de los distintos índices bióticos, y todos estos son ajustes que vienen del ABI y del BMWP/Col; incluso algunos no precisan de 'editing' para tener un NER (μ) adecuado (> 60 %). Con estos datos resulta fácil discernir que el índice biótico óptimo para la CRP es una combinación del ABI + BMWP/Col. (ABI para aplicarse en zonas > 2000 m.s.n.m. y BMWP/Col en lugares < 2000 m.s.n.m.).

Finalmente, se detallan (para los modelos elegidos como óptimos; Tabla 8) los resultados de las herramientas de MATLAB que muestran el FSS, las variables y su frecuencia de uso en la construcción del modelo así como el NER obtenido durante cada una de las 100 corridas realizadas en el análisis.



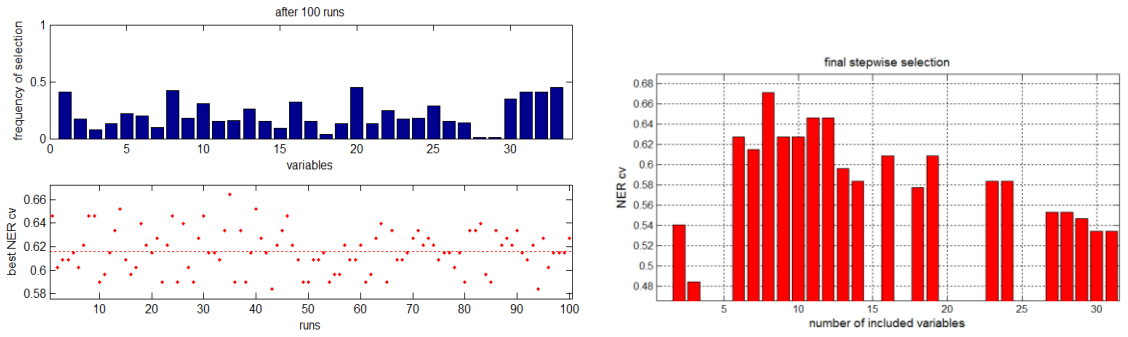
(a) (b)

Figura 23. Índice biótico = ABI con sus datos sin 'editing'. (a) Valores de los NER obtenidos a lo largo de las 100 corridas efectuadas así como la frecuencia de uso de cada una de las 33 variables en el modelo. (b) Final stepwise selection.



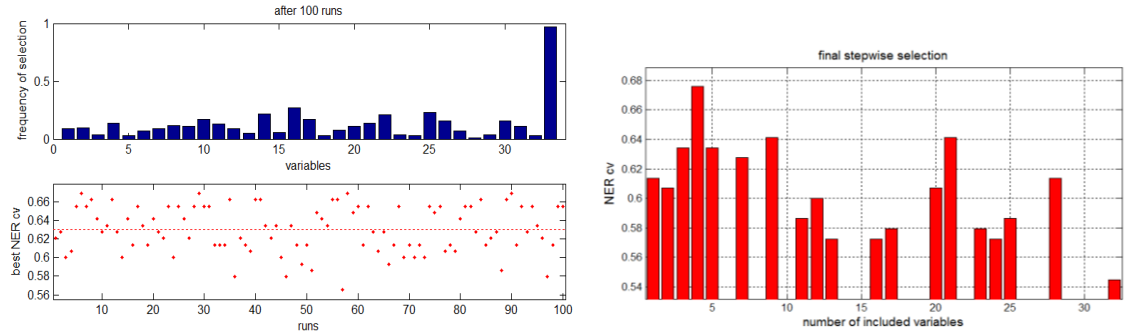
(a) (b)

Figura 24. Índice biótico = BMWP/Col con sus datos sin 'editing' y aplicado solo a los puntos de muestreo ubicados bajo los 2000 m.s.n.m. (a) Valores de los NER obtenidos a lo largo de las 100 corridas efectuadas así como la frecuencia de uso de cada una de las 32 variables en el modelo. (b) Final stepwise selection.



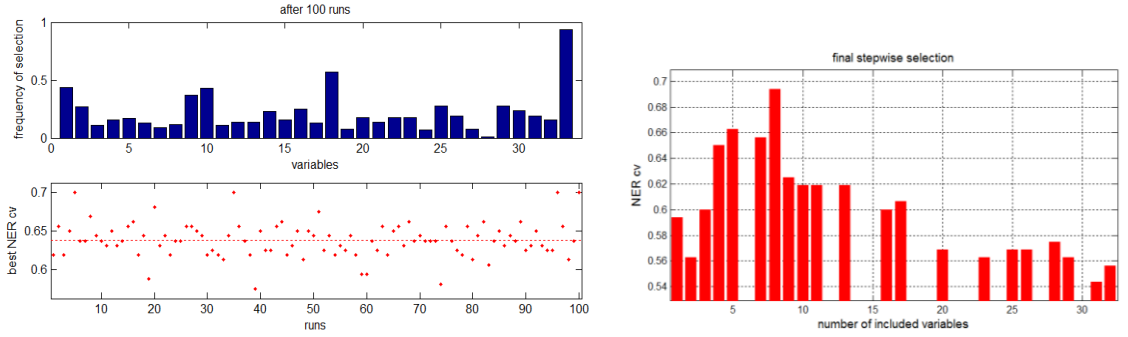
(a) (b)

Figura 25. Índice biótico = ABI + BMWP/Col con sus datos luego del primer proceso de 'editing'. (a) Valores de los NER obtenidos a lo largo de las 100 corridas así como la frecuencia de uso de cada una de las 33 variables en el modelo. (b) Final stepwise selection.



(a) (b)

Figura 26. Índice biótico = ABI con sus datos luego del primer proceso de 'editing'. (a) Valores de los NER obtenidos a lo largo de las 100 corridas así como la frecuencia de uso de cada una de las 33 variables en el modelo. (b) Final stepwise selection.



(a) (b)

Figura 27. Índice biótico = BMWP/Col con sus datos luego del primer proceso de 'editing' y aplicado a toda la CRP. (a) Valores de los NER obtenidos a lo largo de las 100 corridas efectuadas así como la frecuencia de uso de cada una de las 33 variables en el modelo. (b) Final stepwise selection.

Nota: En la Fig. 24 en el gráfico de la izquierda (a) hay 32 variables en lugar de 33 como en el resto de casos. Esto obedece a que en ese modelo específico el BMWP/Col fue calculado solo para la parte baja de la CRP (< 2000 m.s.n.m.) y los valores de la variable Ni siempre fueron cero para este grupo de estaciones, por ende, esta variable fue eliminada del análisis de clasificación.

Tabla 9. Para efectos de interpretación de las figuras 23 a la 27 se detalla el código de etiqueta para cada una de las variables. Para el caso de la Fig. 24, el Al pasa a ser la variable 26 y las restantes corren a partir de esa numeración.

Variable	Código
Orden (Shreve)	1
m.s.n.m.	2
Pendiente	3
Col - T	4
Col - F	5
Ca	6
NO_3	7
NH_4	8
$CaCO_3$	9
Alcalinidad	10
UTN	11
DBO	12
DQO	13
TS	14
Na	15
Fe	16
Cd	17
Mg	18
K	19
Pb	20
CL-	21
F-	22
P	23
PO_4	24
Cu	25
Ni	26
Al	27
pH	28
T (°C)	29
EC	30
OD	31
%_SAT_OD	32
IHF - EPA	33

Finalmente, un resultado interesante se presenta para los modelos construidos en la primera etapa de 'editing' (principalmente) (Tabla 6), pues estos son los que en conjunto poseen el menor número de variables seleccionadas (FSS) por los GAs para llevar a efecto la optimización del *k* – NN (comparar los FSS de la Tabla 6 con los dados en las Tablas 5 y 7). Esta tendencia en aras del 'principio de parsimonia' (Navaja de Ockham)[9] sugiere que ciertamente los modelos construidos a partir de una primera etapa de 'editing' (sobre todo para el ABI y BMWP/Col; tienen mayor NER) son los mejores. En este mismo contexto, resulta lógico suponer que no solo el 'principio de parsimonia' como tal pesa para lograr conclusiones, pero en efecto es un criterio más que suma en pro de que lo modelos generados con el ABI y el BMWP/Col son los óptimos.

[9] Si un fenómeno puede explicarse sin suponer hipótesis alguna, no hay motivo para suponerla. Es decir, siempre debe optarse por una explicación en términos del menor número posible de causas, factores o variables (Kuczera & Mroczkowski, 1998)

8.2. Para la segunda pregunta científica planteada

Una vez que se determinó que la combinación ABI + BMWP/Col fue la variable de respuesta biológica óptima para la CRP, esta fijó y condicionó tres grupos compuestos por las estaciones de muestreo con sus réplicas y las variables que las describen (físico - químicas, microbiológicas y geomorfológicas). Con esta marcha metodológica lo que se esperó fue el determinar qué parámetros del tipo descriptivos explican con mayor peso a los tres grupos preestablecidos y condicionados por los insectos acuáticos (marco conceptual = EH). De tal modo el PCA se llevó a cabo sobre estas tres clases a través de la opción "between group" del software PAST 2.17, para ello los objetos (estaciones de muestreo más sus réplicas) fueron rotulados por tres colores (verde, amarillo y rojo) que corresponden a cada clase predeterminada por los macrozoobentos (Tabla 4). Permitiendo con este proceso analizar con el PCA los grupos etiquetados *a priori* mas no las estaciones de muestreo (Hammer et al., 2001). Se aclara que evidentemente en la matriz sobre la cual el PCA se llevó a cabo no estaba implícita de forma directa la variable correspondiente al índice biótico, pero sí de forma indirecta por los grupos de colores que corresponden a las clases del índice biótico (Fig. 28).

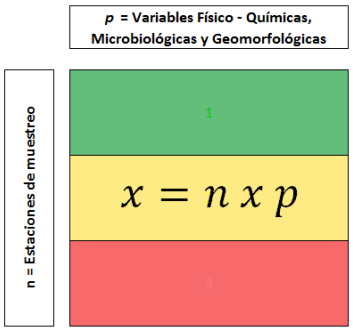


Figura 28. Esquema de la matriz sobre la cual el PCA se llevó a efecto.

De forma concluyente para el PCA llevado a cabo, los resultados que son los óptimos esperados cuando se realiza un análisis de esta naturaleza, la mayor cantidad de varianza explicada en pocos componentes.

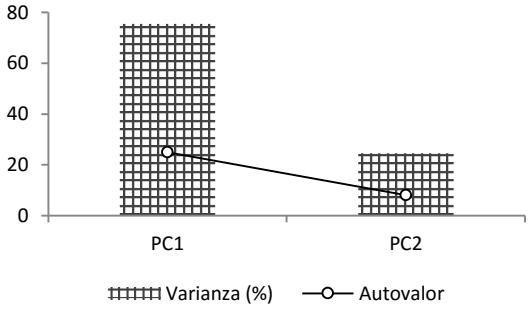


Figura 29. Resultados del PCA.

Tabla 10. Resultados del PCA.

PC / Vector	Autovalor	Varianza (%)
PC1	24,9	75,5
PC2	8,1	24,53

Con estos resultados se evidencia en primer lugar que los tres grupos de estaciones analizadas y condicionados por las tres clases de calidad biótica tienen para un análisis de ordenamiento como el PCA, sentido y explicación en su distribución. Sin embargo, en el Scoreplot del PCA (Fig. 30) no se evidencia una diferenciación clara de los tres grupos pues hay una marcada superposición de puntos. No obstante, si estos se muestran por separado en un Scoreplot individual por clase (Fig. 31, Fig. 32 y Fig. 33) es clara la tendencia del PC1 para explicar mayoritariamente la varianza de los objetos así como la existencia de un patrón de izquierda (PC1-) a derecha (PC2+) que responde a un gradiente de menor a mayor contaminación (Fig. 31, Fig. 32 y Fig. 33).

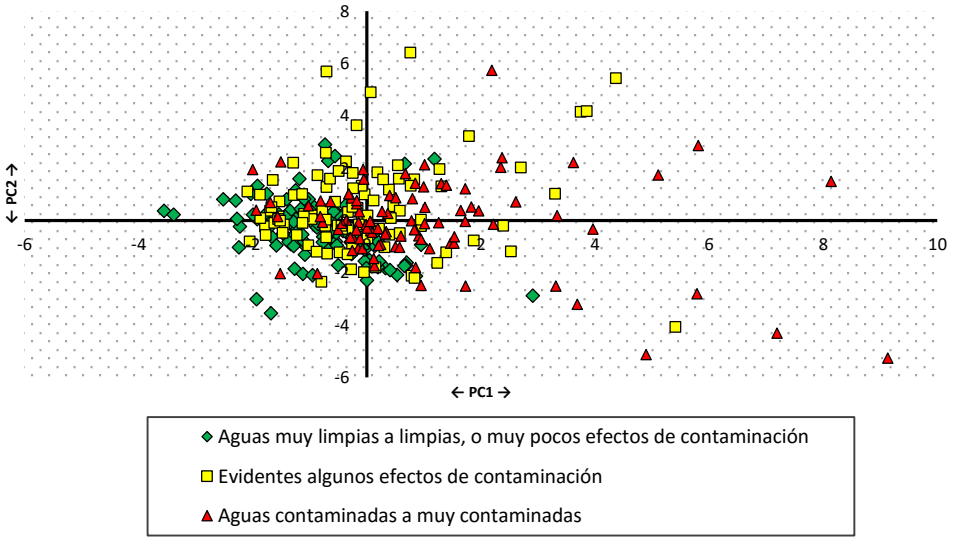


Figura 30. Scoreplot de los tres grupos de estaciones de muestreo determinados por el ABI + BMWP/Col.

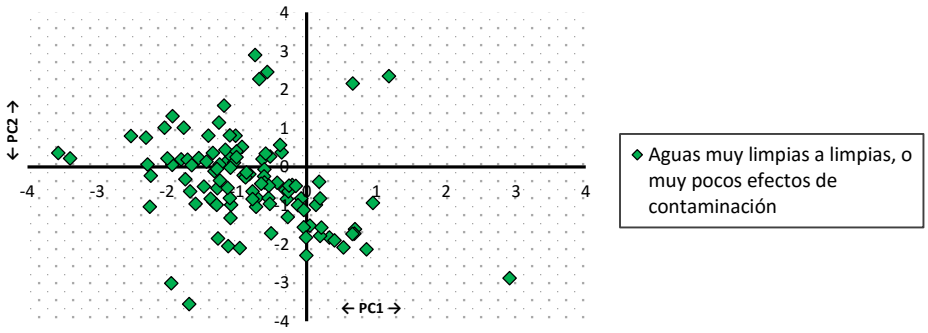


Figura 31. Scoreplot para las estaciones que fueron condicionadas a la clase 1 de calidad biótica.

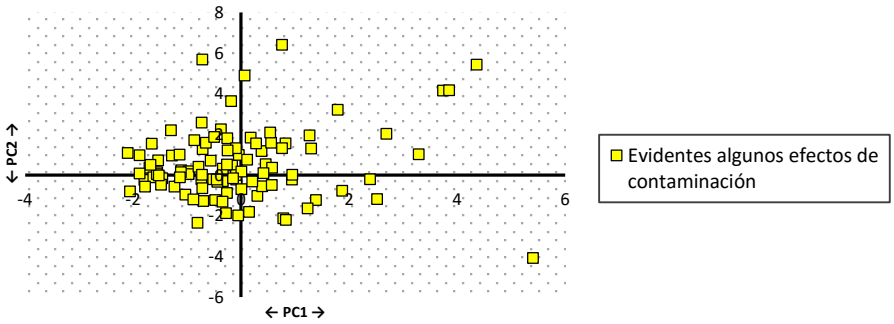


Figura 32. Scoreplot para las estaciones que fueron condicionadas a la clase 2 de calidad biótica.

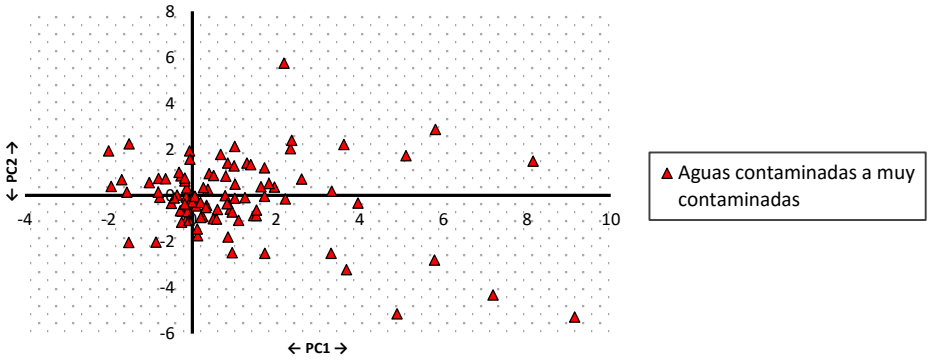


Figura 33. Scoreplot para las estaciones que fueron condicionadas a la clase 3 de calidad biótica.

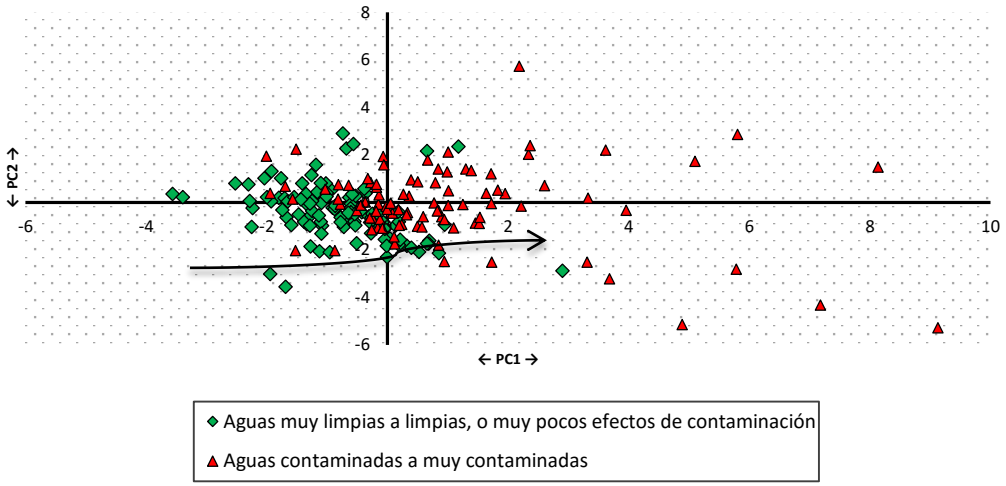


Figura 34. Scoreplot para las estaciones que fueron condicionadas a las clases 1 y 3 de calidad biótica.

Las estaciones de la Clase I están en su mayoría influenciadas por el PC1 – mientras que las Clase III por el PC1+. Objetos que son de la Clase II podrían ser considerados como estaciones de transición entre las clases I y II pues se dispersan sin un patrón claro en los espacios de coordenadas de PC1+ y PC1-. Resta de todo este análisis resaltar el efecto antagónico que existe entre las clases I y II. El criterio de peso > 0,2 (Bere & Tundisi, 2011) fue clave en establecer que variables son las que explican la variabilidad de las clases bióticas de calidad de agua. Los cuatro primeros componentes y los pesos de cada variable hacia ellos fueron los que se tomaron en cuenta para llevar a cabo este análisis aunque siempre se debe prestar especial atención que el PC3 y PC4 no son tan determinantes para explicar la variabilidad del sistema estudiado.

Tabla 11. Peso de las variables originales generados para los cuatro primeros PC.

Variable	Pesos > 0,2 / PC	PC1	PC2	PC3	PC4
Col - F	1	0,2003	-0,0096	0,0272	-0,0406
IHF - EPA	1	-0,2003	-0,0068	0,1425	0,0390
UNT	2	0,1426	-0,2469	0,0514	-0,0182
Alcalinidad	2	0,1359	0,2584	-0,0159	-0,0393
NH_4	2	0,1356	0,2588	-0,0308	-0,0391
Cu	2	0,1154	0,2874	-0,0014	-0,0367
Cd	2	0,0503	0,3403	-0,0289	-0,0256
Al	2	0,0014	0,3515	0,0136	-0,0168
Mg	2	-0,0181	-0,3501	-0,1079	0,0214
%_Sat_OD	2	-0,1190	0,2828	-0,0364	0,0115
Ni	2	-0,1573	-0,2178	-0,0067	0,0421
Temperatura	3	0,1970	0,0647	0,2056	-0,0459
DBO	3	0,1964	-0,0699	-0,2117	-0,0337
pH	3	0,1673	-0,1936	0,3713	-0,0302
OD	3	-0,1721	0,1800	0,6859	0,0170
Pendiente (%)	3	-0,1844	0,1375	-0,3965	0,0366
m.s.n.m.	3	-0,1845	-0,1371	0,2911	0,0398
Orden (Shreve)	4	0,1987	0,0454	0,0137	0,9789
CL-	Pesos < 0,2	0,1997	0,0289	0,0509	-0,0426
EC	Pesos < 0,2	0,1990	-0,0411	0,0436	-0,0391
Na	Pesos < 0,2	0,1986	-0,0472	-0,0328	-0,0377
K	Pesos < 0,2	0,1983	0,0504	0,0365	-0,0431
F-	Pesos < 0,2	0,1977	-0,0571	0,0337	-0,0380
Col - T	Pesos < 0,2	0,1968	-0,0660	-0,0035	-0,0368
Ca	Pesos < 0,2	0,1955	-0,0767	-0,0108	-0,0360
NO3-	Pesos < 0,2	0,1936	-0,0906	0,0092	-0,0352
Dureza Total	Pesos < 0,2	0,1923	-0,0990	-0,0023	-0,0344
DQO	Pesos < 0,2	0,1922	-0,0991	-0,0582	-0,0336
TS	Pesos < 0,2	0,1893	0,1150	0,0273	-0,0442
Pb	Pesos < 0,2	0,1687	0,1896	0,0217	-0,0434
PO_4	Pesos < 0,2	-0,1932	0,0935	-0,0376	0,0354
Fe	Pesos < 0,2	-0,1955	-0,0767	-0,0133	0,0434
P	Pesos < 0,2	-0,1986	-0,0467	-0,0956	0,0438

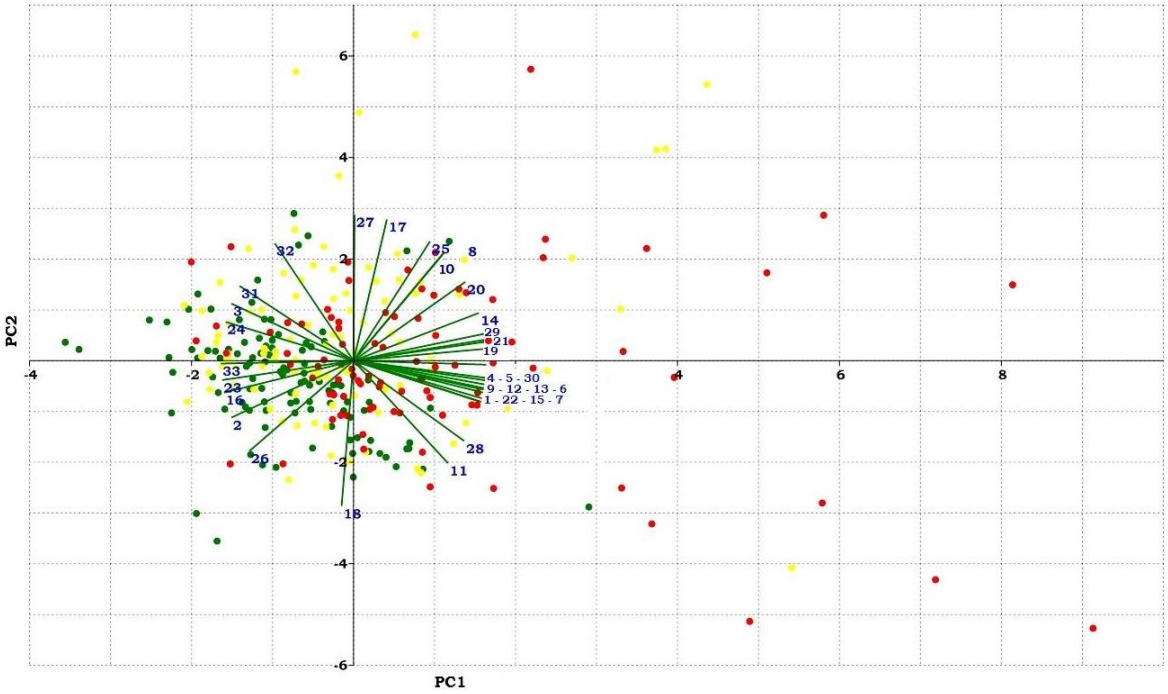


Figura 35. Biplot del PCA. (Para la interpretación del código numérico en azul que corresponde a cada una de las 33 variables utilizar la Tabla 9).

Tanto en la Tabla 11 como en la Fig. 35 se puede observar el peso de las variables originales en las nuevas PC. En la Tabla 11 los valores resaltados con color naranja son los pesos > 0,2 y por ende son los que bajo el PCA describen mayoritariamente el sistema hídrico y su calidad biótica de agua respecto de las variables descriptoras. Son 18 variables de 33 las que el PCA elige como más importantes para explicar uno u otro de los tres grupos (indirectamente = clases bióticas).

No obstante, algunas variables que no cumplen el criterio de peso > 0,2 en el PC1 están muy próximas a consumar esta condición, de modo que aunque fuera del contexto y restricciones asumidos para nuestro PCA se entiende que son relativamente importantes. Además, de una u otra forma estas variables con peso < 0,2 en el PC1 están ligadas a las que son > 0,2 por ende es lógico comprender lo que el PCA proyecta. Por ejemplo Col – F con Col – T, DBO con DQO, Mg con la Dureza Total, entre otros.

De forma más clara en el Biplot (Fig. 35) se proyecta que las estaciones de la clase más limpia (verdes) son influenciadas mayoritariamente por valoraciones elevadas de la calidad de hábitat (IHF - EPA) (33) y niveles adecuados de oxígeno (31); aunque también el níquel (26) condiciona a estos sitios. La altura (m.s.n.m.) (2) y la pendiente (%) (3) igualmente están positivamente ligadas con la Clase I. Por el contrario, sitios representados por la Clase III (aguas contaminadas o muy contaminadas) están claramente caracterizados por pobres coberturas de bosques riparios y una correspondiente homegenización del lecho (33), fuerte presencia de coliformes fecales (5), alta DBO (12) importantes niveles de nitrógeno amoniacal (8) y turbidez (11), pH altos, (28), temperaturas elelvadas (29) y en parte por el orden del río (1), principalmente.

8.2.1. Validación del PCA a través de Regresiones Múltiples

Para adoptar con confianza los resultados obtenidos a través del PCA la fiabilidad científica del mismo debe de ser validada a través del uso de otros métodos independientes, y una forma de lograr este objetivo, es comparar los datos de calidad del agua con y sin las variables que el PCA distingue como 'no principales' (Ouyang, 2005). En este estudio, tres casos se desarrollaron para las comparaciones. En el primero se utilizaron todos los datos de las variables descriptoras para construir un modelo de regresiones múltiples (MR) (Fig. 14) siendo la variable dependiente el índice biótico (ABI + BMWP/Col). En el segundo caso el modelo de MR se llevó a cabo solo con las 18 variables que el PCA identificó como 'importantes' (peso > 0,2) (variables independientes) para explicar el ABI + BMWP/Col (variable dependiente). En un tercer y último caso las variables independientes fueron 15 y correspondieron a las consideradas como 'no importantes' por el PCA. La variable dependiente fue la misma.

Estos tres casos se compararon para determinar si la exclusión o no de datos de las variables 'no principales' influyó en las relaciones de regresión múltiple.

Tabla 12. Parámetros de la MR en los tres casos llevados a cabo para la validación del PCA.

Parámetros	Caso 1	Caso 2	Caso 3
No. Variables Independientes	33	18	15
R^2 (%)	40,5	34,6	13,14
R^2 - ajustado (%)	33,1	30,4	8,6
Error estándar	23,7	24,2	27,7
Error absoluto medio	17,83	19,02	21,9
Durbin-Watson (P = 0,0000)	0,8	0,7	0,3
Autocorrelación residual en Lag 1	0,6	0,64	0,84

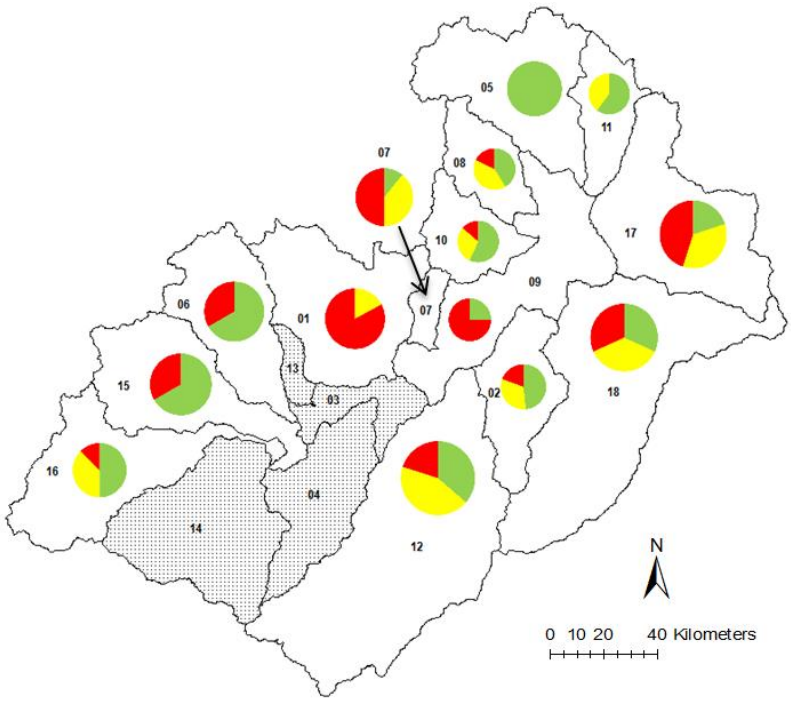
El estadístico R^2 indica el porcentaje de variabilidad que explica el modelo para el ABI + BMWP/Col, sin embargo el estadístico R^2 - ajustado (%) es más ventajoso para comparar modelos con diferentes números de variables independientes como ocurre en el presente caso. El error estándar de la estimación muestra la desviación típica de los residuos. El error absoluto medio (MAE) es el valor medio de los residuos. El estadístico Durbin-Watson (DW) examina los residuos para determinar si hay alguna correlación significativa basada en el orden en el que se han introducido los datos.

Con estos parámetros de base para los modelos de MR se evidencia que el R^2 - ajustado varía de forma insignificante cuando las 15 variables calificadas como 'no importantes' se extraen del modelo de MR (Tabla 12). Esto mecánicamente valida los resultados del PCA y a su vez el criterio de que variables con pesos menores a 0,2 (Bere & Tundisi, 2011) son 'menos importantes' en explicar la variabilidad de la calidad biótica de agua en la CRP.

El tercer caso fue llevado a cabo como una prueba de tipo antagónica y en efecto se pudo demostrar que las variables descriptoras seleccionadas son 'poco importantes' para explicar al ABI + BMWP/Col y su variabilidad (Tabla 12).

Este método de validación del PCA manifiesta una categórica eficiencia y apunta a que varios parámetros medidos a lo largo de la serie temporal considerada son excusables y por tanto, una simplificación (fuerte) del modelo inicial de 33 variables descriptoras es posible.

Una vez que se caracterizaron las clases I y III a través de las principales variables descriptoras, un resultado interesante fue el análisis a nivel espacial de la CRP. Para ello un análisis del tipo gráfico / descriptivo se llevó a cabo a nivel de subcuencas hídricas dentro de la CRP.



Mapa 6. Incidencia porcentual de las clases de calidad de agua en las subcuencas de la CRP (Con un fondo degradado por puntos están los subsistemas que no fueron monitoreados).

Subcuencas como Burgay (01), Magdalena (07), Paute (09) y en cierta medida la Zona Baja del Paute (17) son sitios en donde la calidad de las aguas es bastante mala. Para estas zonas en su gran mayoría todas las estaciones de muestreo a lo largo de la serie temporal incurrieron en categorías de Clase III (aguas contaminadas a muy contaminadas). Al contrario, subsistemas como Púlpito (11), Juval (05), Mazar (10), Pindilig (10) y en cierta medida, Yanuncay (16), Machángara (06) y Tomebamba (15) poseen aguas con buenas condiciones para los macrozoobentos siendo la mayoría de sus estaciones clasificadas dentro de la Clase 1 (aguas muy limpias a limpias, o muy pocos efectos de contaminación).

En una situación alternativa se encuentran las aguas de las zonas del Santa Bárbara (12), Collay (02) y Negro (18). Estos puntos presentan un cierto aumento de las estaciones Clase II (evidentes algunos efectos de contaminación), de tal modo, podrían ser considerados como cuerpos de agua en donde hay factores que causan estrés pero no son lo suficientemente fuertes como perturbar drásticamente a la comunidad de macrozoobentos.

8.3. Redistribución de las clases de índices bióticos

En pruebas previas llevadas a cabo para cada modelo generado a través de los distintos índices bióticos y sus variantes, se evidenció que estadísticamente tanto para el método de clasificación (*k* – NN + GAs) como en el de ordenamiento (PCA) hay un proceso de optimización de las clases bióticas (según indicadores de rendimiento estadístico) y su capacidad de respuesta biológica cuando estas son tres en lugar de las cinco estándar.

Estas cinco clases patrón al parecer dictaminan que las fronteras o bordes de cada una tengan un perfil muy distintivo y hasta cierto punto excepcional, por ende el reconocimiento de patrones (correcta asignación de objetos a las clases *a priori*) que se pretende con el *k* – NN + GAs es de una calidad o pureza deficientes. Se considera que una explicación para esto es que la repuesta biológica dada por los macrozoobentos, si bien es muy adecuada, no tiene un carácter lineal, mucho menos en un complejo sistema con 33 variables descriptoras (varias de éstas independientes entre sí).

Con estos antecedentes, formas de determinar clases bióticas más apropiadas (menos robustas) fueron analizadas y resultó que una reestructura de 5 a 3 clases fue motivo de una mejora evidente en el desempeño del *k* – NN + GAs. Estas 3 clases se extrajeron de los índices bióticos a través del cálculo de los percentiles 33,33 % y 66,66 %.

Tabla 13. Modelo de clasificación llevado a cabo con el índice ABI + BMWP/Col con cinco clases bióticas. Sin proceso de 'editing'.

Índice Biótico	FSS	Modelo de Clasificación (*k* - NN + GAs)								NER (μ)	NER (SD)
		Peso de Variables									
ABI+BMWP/Col (con las 5 clases estándar)	7	Variables Seleccionadas	OD	IHF - EPA	EC	Ni	Fe	Pb	Shreve	0.51	0.03
		NER	0.40	0.50	0.43	0.43	0.50	0.50	0.51		
		Frecuencia de Selección	45	42	41	35	32	31	29		

El ABI + BMWP/Col es la variable de respuesta biológica que más se ajusta para un modelo de clasificación, empero al observar la Tabla 13 en la cual esta misma medida biótica es expresada en 5 clases, disminuye su NER significativamente si se compara con datos de la tabla siguiente (ABI + BMWP/Col con tres clases bióticas):

Tabla 14. Modelo de clasificación llevado a cabo con el índice ABI + BMWP/Col con tres clases bióticas determinadas por los percentiles 33,33 % y 66,66 %. Sin proceso de 'editing'.

Índice Biótico	FSS	Modelo de Clasificación (*k* - NN + GAs)								NER (μ)	NER (SD)
		Peso de Variables									
ABI+BMWP/Col (con las 3 clases determinadas por percentiles)	8	Variables Seleccionadas	Pb	IHF - EPA	NH4+	Shreve	OD	T(°C)	EC	Fe	0,63	0,06
		NER	0,54	0,5	0,63	0,6	0,7	0,63	0,63	0,65		
		Frecuencia de Selección	45	42	41	35	32	31	29	26		

En un contexto análogo, resultados para el PCA llevado a cabo sobre las clases de datos del tipo descriptores y clasificados por el ABI + BMWP/Col en cinco grupos estándar, revelan que la varianza explicada disminuye significativamente en el PC1 y PC2 si se compara cuando el ABI + BMWP/Col es expresado en tres clases (percentiles 33,33 % y 66,66 %) (Tabla 15).

Tabla 15. Resultados del PCA realizado entre 3 y 5 grupos de datos descriptores.

PC / Vector	ABI + BMWP/Col expresado en 5 clases		ABI + BMWP/Col expresado en 3 clases	
	Autovalor	Varianza (%)	Autovalor	Varianza (%)
1	16,4	49,6	24,9	75,5
2	9,7	29,4	8,1	24,5
3	4,8	14,6	-	-
4	2,1	6,5	-	-

Se intuye que las razones para el aumento del rendimiento estadístico cuando las clases cambian de cinco a tres radica en el hecho de que disminuye la robustez de las mismas y se evita la búsqueda de patrones netamente lineales entre las variables de repuesta biológica y las descriptoras.

8.4. Los macroinvertebrados bentónicos

Un total de 53452 individuos de macroinvertebrados bentónicos han sido colectados. Taxonómicamente estos se corresponden con 18 órdenes y 65 familias. Las órdenes más representativas son los Efemerópteras (63,7 %), Díptera (11,9 %), Trichoptera (8,5 %), Coleóptera (6,3 %) y Anélida (5,2 %), las restantes taxas son minoría en la muestra global colectada. En lo referente a las familias la abundancia total mayoritariamente sigue recayendo en taxones similares como Beatidae (Efemeróptera) (55,9 %; Fig. 36).

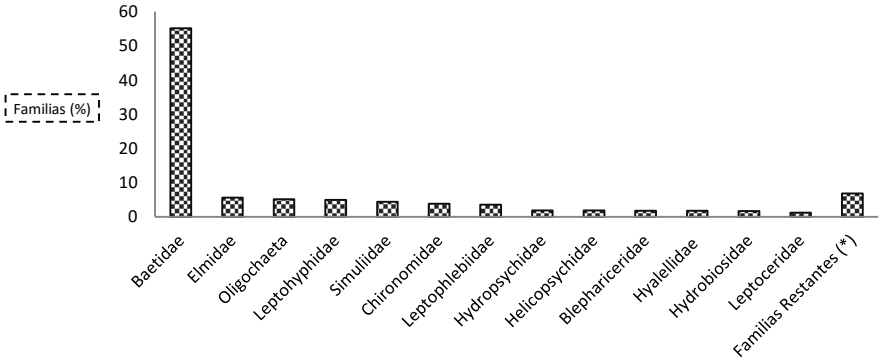


Figura 36. Familias de macrozoobentos más representativas durante los muestreos. (*) Son las 53 familias restantes (8,14 %).

9. DISCUSIÓN

No es insólito que el ABI más el BMWP/Col aplicados en las partes altas (> 2000 m.s.n.m.) y bajas (< 2000 m.s.n.m.) respectivamente sean la mejor respuesta biológica de todas las analizadas. En la CRP un 92,92 % de la superficie corresponden a zonas mayores a los 2000 m.s.n.m. y como consecuencia de esto la mayoría de cómputos bióticos incumbieron para el ABI, siendo esta, una medida calibrada específicamente para zonas andinas y con esto su calidad de aproximación en bioindicación es evidentemente mayor. Ríos-Touma et al. (2006) han desarrollado a través de gabinete y considerando los estudios de autoecología de los taxones de macrozoobentos andinos, scores adecuados que reflejen la calidad de las aguas superficiales. De hecho nuevos ajustes se han efectuado recientemente (Ríos-Touma et al., 2014) lo que en cierto modo valida el esfuerzo y por ende la confiabilidad de los trabajos realizados en pro mejora de la calibración del ABI.

Por otro lado, es evidente que en la CRP los scores para ciertos taxones no van a funcionar correctamente si se pretende que estos sean generalistas a toda la cuenca porque la heterogeneidad del sistema hídrico no lo permite (este problema fácilmente puede ser replicado a todo el Ecuador).

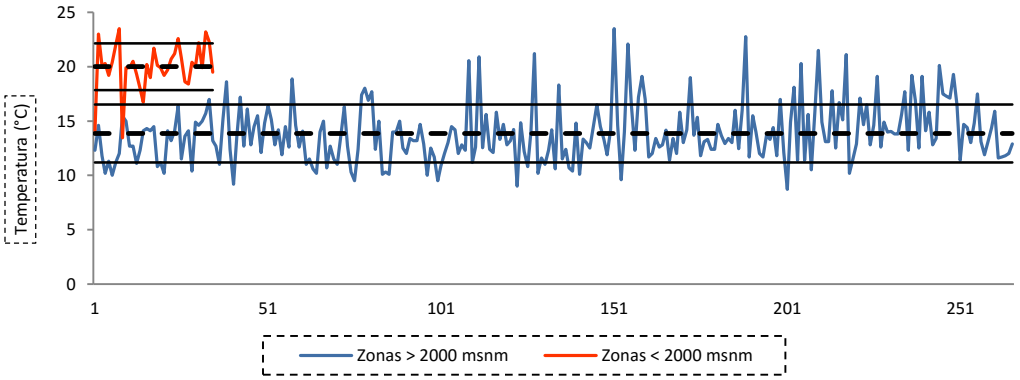


Figura 37. Según un diferencial de altura se especifican las temperaturas del agua medidas en este estudio. Líneas entre cortadas = promedios; líneas sólidas = SD.

Si se considera a la Fig. 37, es natural pensar que muchas más variables del tipo geomorfológicas van a determinar la presencia de ciertas taxas exclusivas de zonas bajas y otras de partes altas. De tal modo que el uso de un solo índice biótico (de los disponibles hasta el momento) podría suponer sesgos importantes en bioindicación para la CRP.

Por estas razones es necesario discutir aspectos de la familia Beatidae pues es un taxón que luego del análisis de datos efectuado plantea necesidades de reajustes a sus scores para representar adecuadamente sucesos de estrés en los ríos y quebradas de la CRP.

Para Ecuador un total de 10 especies de Beatidae han sido descritas hasta el momento: *Callibaetis nigrivenosus* (Banks, 1918), *Andesiops peruvianus* (Ulmer, 1920), *Callibaetis camposi* (Navás, 1930), *Baetodes levis* (Mayo, 1968), *Baetodes spinae* (Mayo, 1968), *Prebaetodes sitesi* (Lugo-Ortiz & Mccafferty, 1996), *Mayobaetis ellenae* (Mayo, 1973), *Nanomis galera* (Lugo-Ortiz & Mccafferty, 1999), *Varipes lasiobrachius* (Lugo-Ortiz & Mccafferty, 1998) y *Americabaetis robacki* (Lugo-Ortiz & Mccafferty, 1994), en Domínguez et al. (2006). Empero, no hace falta ahondar respecto a que la riqueza de esta familia está muy por debajo de ser representada de forma confiable en el país. En contexto, durante los años que abarco el presente monitoreo, los géneros registrados de Beatidae en la CRP fueron: *Andesiops*, *Americabaetis*, *Baetodes*, *Callibaetis* y *Mayobaetis*. Las especies no fueron determinadas pues se necesitó la asociación ninfa – adulto y el foco de este estudio no era esencialmente taxonómico. Los hallazgos sistemáticos / geográficos registrados en este estudio evidencian que Beatidae, al agrupar a géneros como *Baetodes*, no puede ser una familia para la cual se pretenda que un score asignado sea generalista y apropiado a la vez. El género *Baetodes* estuvo presente con altísimas abundancias en zonas como el Burgay y Magdalena que tienen una fuerte contaminación por desechos orgánicos (principalmente aguas negras), por el contrario, individuos de *Mayobaetis*, *Andesiops* y *Americabaetis* tendencialmente prefirieron sitios limpios como aguas de las subcuencas de Mazar, Pindilig, Púlpito o Negro aunque también estuvieron presentes en zonas perturbadas pero en números mucho menores. Los miembros de *Baetodes* corresponde con un 25,13 % del total de la muestra de macrozoobentos, de modo que en un análisis macro, estos no conciernen o se explican según un gradiente de contaminación como el que se observa en la CRP, es decir, para la zona objeto de estudio no son tan sensibles a efectos de estrés como se cree. Tendencias similares de altísimas abundancias de *Baetodes* fueron reportadas en Brasil (Buss & Salles, 2007). Los autores describen una gran preferencia de estos individuos en sustratos rocosos y señalan que presentan adaptaciones morfológicas para resistir el estrés hidráulico. Para nuestro caso la tendencia a una explosión poblacional por parte de *Baetodes* es mucho más referida en las zonas altas mayores a los 2500 m.s.n.m. y principalmente en la subcuenca del Burgay, donde

factores como elevados conteos para coliformes fecales así como una muy baja valoración de la calidad del hábitat fluvial fueron características primarias.

Por otro lado, todos estos elementos de alta variabilidad (± SD) para los datos de abundancias de macrozoobentos jugaron un rol importante en la selección de la medida biótica óptima. Índices como el que aquí se propone (CRP Index) se vieron seriamente sesgados al ponderar la abundancia como un factor ecológicamente positivo cuando en el caso del Burgay ésta al parecer sigue a algún factor de estrés hidroquímico. EPT también fue una medida que no puede considerarse del todo adecuada para la CRP pues algunas efemerópteras (principalmente Beatidae, *Baetodes*) están presentes ampliamente en zonas que indudablemente están degradadas.

Un resultado interesante lo proyecta el EPT / EPT + OCH (Bonada et al., 2006). En los modelos de clasificación generados con este índice el NER fue bastante bueno [NER 0,6452 (μ)] aunque para ello hizo falta un proceso de 'editing' algo exagerado. Sin embargo, el EPT / EPT + OCH se diseñó para reflejar detrimentos del caudal que indican un avance de periodos más secos (Feminella, 1996 & Boulton, 2003; Chaves et al., 2008), de tal forma que hay ríos y especialmente quebradas dentro de la CRP que son más susceptibles a las disminuciones de flujo de agua. Es por eso que para sitios de esta naturaleza el EPT / EPT + OCH podría proporcionar resultados adecuados.

En lo referente a las variables que se distinguen mayormente para explicar la naturaleza de la calidad biótica del agua en la CRP, se observa que hay una significativa correspondencia entre los resultados del *k* – NN + GAs y el PCA para determinar los descriptores más importantes. Aunque el *k* – NN + GAs no se lo empleó con estos propósitos, esto constituye un aporte importante, pues a pesar que ambos métodos proceden de bases estadísticas distintas (clasificación y ordenamiento respectivamente) su coincidencia demuestra que en efecto ciertas variables físico – químicas, microbiológicas y geomorfológicas son más trascendentales para explicar las clases bióticas asignadas por el ABI + BMWP/Col.

Las variables más significativas para explicar a los macrozoobentos y sus tres clases dadas por el ABI + BMWP/Col son en primera instancia, la calidad del hábitat fluvial (IHF – EPA; Fig. 38).

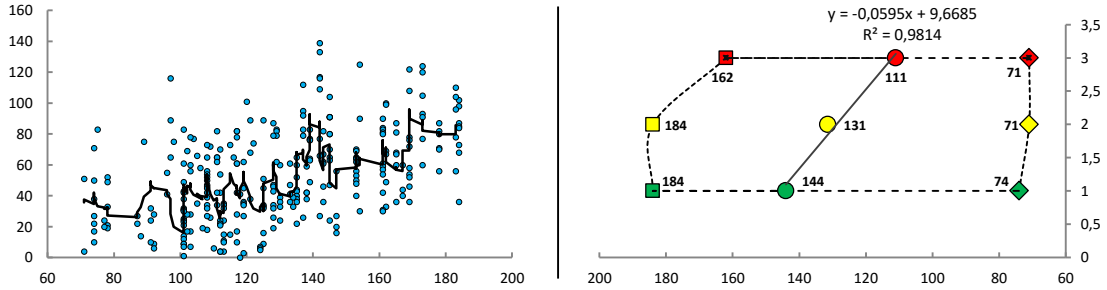
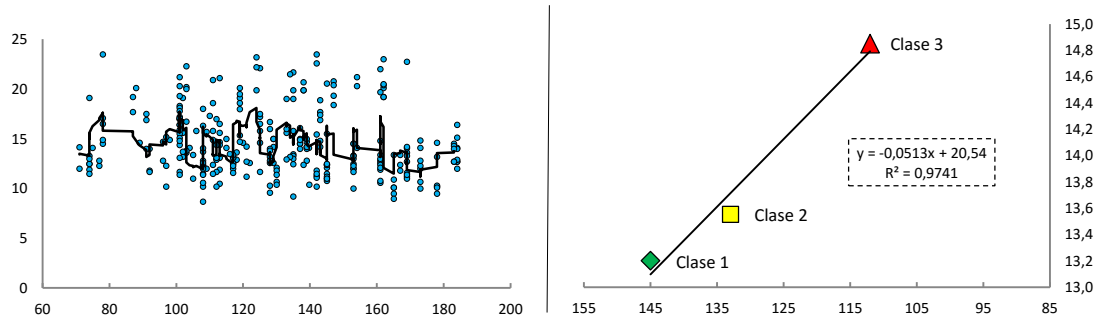


Figura 38. (a) Media móvil del ABI + BMWP/Col respecto del IHF - EPA; (b) modelo lineal de las tres clases dictadas por el ABI + BMWP/Col respecto del IHF – EPA (Cuadrados = máximos; rombos = mínimos; círculos = μ del IHF – EPA para las Clases I, II, III).

Según los datos de este estudio, mientras mejores sean las condiciones del hábitat fluvial, (buena cobertura y estructura de los bosques riparios y mayor heterogeneidad del lecho), las comunidades de macrozoobentos son mejores tanto a nivel de diversidad como de estructura. Así también se presentan taxones indicadores de aguas no contaminadas (Fig. 38) (Clase I) y en sí, la calidad del agua en general es mejor. Los ecosistemas ribereños (que son uno de los condicionantes principales de la estructura del lecho) implican una perspectiva geomorfológica

holística producto de la extensa serie interconectada de biotipos y gradientes ambientales que, con sus comunidades bióticas, constituyen en sí los sistemas fluviales (Ward, 1998). Es por ello la sensibilidad directa observada en los macrozoobentos frente a un declive de los índices de hábitat fluvial. Estrategias de manejo de los bosques riparios son prácticas ya conocidas con el fin de mejorar la calidad de las aguas (Comerford, 1992). Un estudio realizado en los E.E.U.U. demostró que los ecosistemas riparios son capaces de retener entre un 50 % y un 90 % de la carga total de sedimentos de la escorrentía superficial y el nitrógeno total de esta, así como los nitratos de las aguas subterráneas poco profundas (Lowrance et al., 1997). Estos hallazgos se ajustan marcadamente a los nuestros pues las estaciones que se identifican como contaminadas (Clase III) y que tienen bajo IHF – EPA son las que ostentan niveles altos de NH_4 y NO_3 (NO_3 es una variable que bajo el PCA y el *k* –NN + GAs fue calificada como 'poco importante').

Otro resultado interesante en base a los datos de este estudio es que existe una tendencia al aumento de la temperatura del agua en las estaciones Clase III. Se cree que este patrón se explica también (en parte) por los ecosistemas riparios, pues cuando éstos son pobres o nulos, la incidencia de la radiación solar sobre el canal de los ríos y quebradas es directa (Johnson, 2003).



(b) **(b)**

Figura 39. (a) Media móvil para la temperatura en función del IHF – EPA; **(b)** modelo lineal de los promedios de temperatura respecto del IHF – EPA para las tres clases dictadas por el ABI + BMWP/Col.

Resultados similares a los aquí expuestos (Fig. 39) han sido publicados (Beschta et al., 1987), aunque existe en la actualidad una discusión grande y controversial respecto a los factores que determinan la dinámica de la temperatura. Hay numerosos trabajos publicados acerca de los factores de control de temperatura en sistemas lóticos: el rol de la temperatura del aire (Sullivan & Adams, 1991), efectos de sombra en el cauce (Beschta, 1997), flujos de sustrato y de conducción (Webb & Zhang, 1997) y los cambios en las trayectorias de temperatura debido a la longitud (Beschta et al., 1997). Lejos de esta discusión, se piensa que este aumento gradual observado en la temperatura en favor de un gradiente de contaminación y de no 'equilibrio ecohidrológico' es debido a varios factores, pero uno importante para este caso es el de la desprotección riparia.

Una siguiente variable señalada como importante por el PCA (el *k* – NN + GAs no la consideró 'importante') son los coliformes fecales (Col - F). Conteos elevados de este grupo de microorganismos están fuertemente ligados con las estaciones de mala calidad de agua y en sí determinan un gradiente de contaminación orgánica de menor a mayor desde la Clase I a la III (Fig. 40).

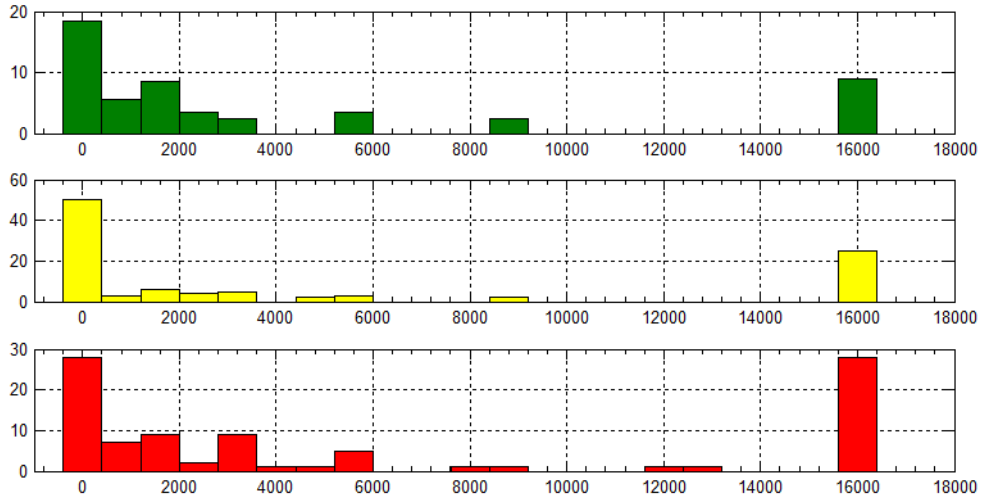


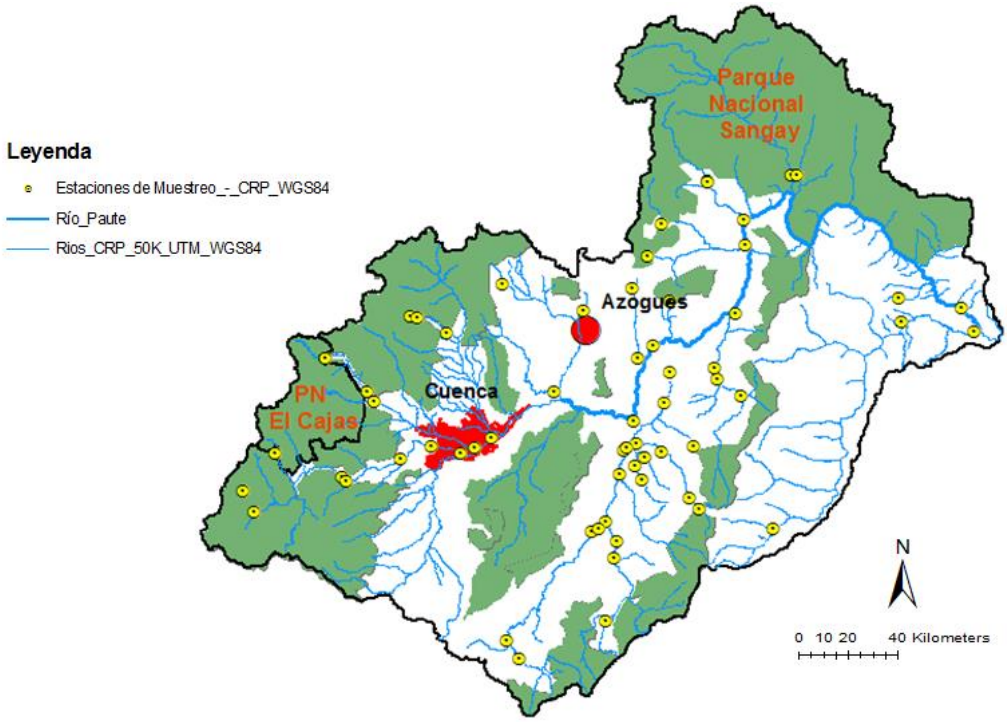
Figura 40. Histogramas de frecuencia para Col – F en las tres clases de datos (Verde = Clase I, Amarillo = Clase II, Rojo = Clase III).

El histograma de la Fig. 40 es claro al mostrar que la mayor frecuencia de sitios con niveles muy elevados de Col – F está predominantemente en la Clase 3.

Los coliformes fecales están presentes sobre todo en las zonas del Burgay y Magdalena y en menor medida en Paute y su parte baja; estas dos primeras son las subcuencas con menores IHF – EPA (Clase III). Pese a esto, es claro que la mala calidad de estos subsistemas satisface a un contexto holístico de circunstancias inadecuadas para el 'equilibrio ecohidrológico' y no solo a pocos factores. Burgay y Magdalena tienen una alta incidencia de prácticas agroindustriales, principalmente de cultivos y ganadería. Igualmente la ciudad de Azogues (33848 habitantes) se ubica en la parte media del Burgay y no posee tratamiento para sus aguas residuales, las cuales son descargadas al final del sistema pocos kilómetros antes de su unión con el Paute (curso principal). Esto explica que la mayor frecuencia de conteos muy elevados de Col – F sean característicos de la Clase 3 (Fig. 40).

A la par, se cree que la contaminación de las aguas para estas dos subcuencas en gran parte radica en el hecho de no poseer adecuados sistemas riparios. La escorrentía superficial transporta a los causes grandes cantidades de los desechos agroindustriales (material fecal, fertilizantes, pesticidas, etc.), además de altas cargas de sedimentos. Asimismo este aporte extra de sedimentos sumado a la NO introducción de materia orgánica producto de los bosques riparios (ramas, hojas, troncos, et.) acarrean la homogenización de los lechos con lo cual disminuye drásticamente la cantidad de nichos para la especializada comunidad de macrozoobentos.

Finalmente, con el cúmulo de antecedentes que se detallan se deduce que IHF – EPA es la mayor condicionante en determinar y explicar la calidad del agua en la CRP. Se observa un dinámico y claro patrón espacial entre las estaciones de muestreo catalogadas como Clase I (aguas muy limpias a limpias, o muy pocos efectos de contaminación) y la presencia de bosques o áreas protegidas o con muy poca antropización (Mapa 7) (el efecto adverso de este patrón es también observado).



Mapa 7. Áreas de bosque y vegetación protegidas y las ciudades principales (en rojo) de la CRP.

En el Mapa 7 se puede ver que zonas como Burgay, Magdalena, Paute y su parte baja son sitios que no poseen bosques protegidos o parques nacionales, por el contrario, subcuencas como Púlpito, Mazar, Pindilig y las zonas de la parte alta de la CRP tienen importantes remanentes de bosques nativos que han sido declarados como zonas a proteger. Este patrón indica la función depuradora que brindan los ecosistemas ribereños sobre los ríos y quebradas de la CRP.

10. CONCLUSIONES Y RECOMENDACIONES

Dados los óptimos valores en los parámetros que miden el rendimiento estadístico de las metodologías empleadas (*k* – NN + GAs y PCA), es lógico que los resultados presentados (selección del ABI + BMWP/Col; y la determinación de algunas 'variables importantes' y otras 'no importantes') sean tomados en cuenta como una fuerte herramienta de gestión en pro mejora del adecuado manejo de los recursos hídricos.

Las marchas multivariantes utilizadas en este estudio prueban ser potentes herramientas para investigaciones bajo el marco conceptual de la EH, por ello su uso y potencialización son recomendados.

Los modelos generados por el *k* – NN + GAs y que se han seleccionado como óptimos (ABI + BMWP/Col), proceden de métodos sensatos dentro de los cuales el 'no abuso' de ciertas técnicas como el 'editing' fueron auto condicionamientos que se plantearon desde un inicio. Los valores de NER no son en extremo óptimos pero son bastante aceptables. Así, el propósito de emplear el *k* – NN + GAs no fue precisamente la modelización en predicción sino elucidar cuál de las distintas medidas bióticas ajusta mejor en un modelo de clasificación. El criterio arbitrariamente elegido acerca que un máximo 15 % de objetos podrían ser eliminados en el 'editing', resulta bastante lógico pues no se forzó a calcular un modelo sobre un conjunto de datos homogenizados.

La determinación de variables importantes bajo el prisma del PCA y el criterio que éstas sean solo las que tengan pesos > 0,2 brindó resultados muy operativos. Dicho criterio resultó ser capaz de sintetizar la variabilidad del sistema satisfactoriamente. Asimismo, el listado de variables para futuras campañas de monitoreo podría ser reducido en ciertos aspectos con significativos ahorros de tiempo y dinero para las entidades gestoras de los recursos hídricos de la CRP.

El PCA es un procedimiento de amplio uso en ciencias ambientales, sin embargo pocos son los trabajos que intentan ganar criterio en términos de validación de resultados a través de este método. El proceso que se ha implementado para asegurar la rigurosidad científica del PCA resulta una simple pero potente herramienta.

Dada la gama de diversidad de los macroinvertebrados bentónicos y al alto nivel de especialización que presentan como comunidad de organismos, se constituyen como una excelente variable de respuesta biológica. Sin embargo, para una correcta bioindicación el nivel de resolución taxonómica para ciertos grupos como los efemerópteros debe de ser perfeccionado. En zonas con una alta heterogeneidad como la existente en la CRP un 'score' como por ejemplo el de Beatidae puede sesgar marcadamente un resultado, para evitar ello se recomienda ahondar en la sistemática y autoecología de especies de la familia Beatidae, principalmente para que las futuras aproximaciones hacia la elaboración de juicios de calidad biótica del agua sean adecuadas.

Aunque no se ha realizado un análisis puntual sobre los datos de macrozoobentos, se puede señalar que los taxones más sensibles a efectos de contaminación son: Perlidae, Gripopterygidae, Psephenidae, Ptilodactylidae, Leptophlebiidae, Calamoceratidae, Glossosomatidae y Xiphocentronidae. Casi son nulos en muestras clasificadas como Clase 3 e incluso algunos, en la Clase 2 ya desaparecen.

El EPT / EPT + OCH fue diseñado para reflejar detrimentos del caudal que indican un avance de periodos más secos, de tal forma, hay ríos y especialmente quebradas dentro de la CRP que son más susceptibles a las disminuciones de flujo de agua. En estos casos es posible el EPT / EPT + OCH podría proporcionar resultados adecuados. Esta herramienta también podría ser valorada como una opción para criterios preliminares en temas de caudal ecológico en ríos o quebradas donde su uso sea permisible.

La variable que más explica condiciones adecuadas de calidad de agua es el IHF – EPA. En la CRP y en general, las áreas ribereñas constituyen una pequeña proporción de la superficie total de la cuenca, sin embargo, su rol en los procesos que determinan un estado u otro de calidad del agua es clave. En este trabajo un protocolo de valoración producido para Norteamérica fue el utilizado, de tal modo, es posible que ciertos aspectos no puedan ser medidos de forma óptima en la CRP. Es por ello que la construcción de metodologías para valorar ecosistemas riparios y la calidad de los lechos estandarizadas y calibradas a las condiciones locales, son urgentes. Con los resultados del presente estudio se crea, para este aspecto puntual, una notable necesidad de investigación.

Asimismo, los datos muestran que existe una tendencia al aumento de la temperatura del agua en las estaciones Clase III. Este patrón es igualmente explicado (en parte) por los ecosistemas riparios pues cuando estos son pobres o nulos la incidencia de la radiación solar sobre el canal de los ríos y quebradas es directa. Se recomienda con el fin de ahondar en esta ciertamente, controversial temática, en futuros estudios también tomar medidas de la temperatura ambiente.

Las clases 1 y 3 fueron las más tendenciales a ser caracterizadas óptimamente por el PCA. De tal modo, la Clase 1 es descrita principalmente por IHF – EPA elevados, niveles de OD adecuados y una pendiente ligeramente pronunciada. Por el contrario la Clase 3 se identifica por altísimos conteos de coliformes fecales, elevados niveles de nitrógeno amoniacal, turbidez marcada y altas temperaturas del agua que se correlacionan con una disminución o ausencia de bosques ribereños y una homogenización del lecho. Valores de pH elevados también son asociados a las Clase 3.

A nivel espacial y de subsistemas hídricos, Burgay, Magdalena, Paute y su parte baja son los peores sitios en términos de calidad de agua. Se asocian directamente con la Clase 3. Mientras que Juval, Púlpito, Pindilig, Mazar, y las subcuencas de la parte alta, adyacentes al Parque Nacional El Cajas, son los sitios asociados con la Clase 1, por ende los que mejor status de calidad de agua y de integridad ecológica presentaron. Siendo este patrón correlacionado directamente con las estrategias de conservación de los bosques nativos que se llevan a cabo en la CRP.

Se conjetura que las razones para el aumento del rendimiento estadístico cuando las clases cambian de cinco a tres, radica en el hecho que disminuye la robusticidad de las mismas y se evita la búsqueda de patrones netamente lineales cuando estos no necesariamente existen. Asimismo, la simpleza al dividir los índices bióticos por medio de los percentiles 33,33 % y 66,66 % resultó ser una estrategia bastante óptima.

Con los análisis efectuados se crea una línea base de calidad de agua para la CRP bastante plausible, en tal virtud a la par de seguir proporcionando diagnósticos adecuados de los sistemas hídricos superficiales, la comunidad involucrada en el manejo y gestión de los recursos hídricos debe de apuntalar a planes que contemplen la restauración de los ecosistemas acuáticos degradados. Una pieza clave en un proceso piloto para el caso puntual de la CRP sería la restauración de los ecosistemas ribereños.

Finalmente, una importante lección para el futuro, radica en el hecho que resultados puntuales generados en latitudes diferentes y distantes y que se usan en ecología acuática o ciencias afines, no necesariamente sirven en otras regiones. Así por ejemplo el EPT es una prueba palpable de esto, pues taxones como Beatidae pueden sesgar de forma severa los resultados de una muestra. Se recomienda que a través de métodos de optimización matemática o estadística como por ejemplo los aquí presentados, se debe tratar de validar el uso formal de protocolos u otras herramientas provenientes de sitios ajenos a la realidad biogeográfica, y socioeconómica de una zona.

11. REFERENCIAS BIBLIOGRÁFICAS

Acosta Rivas, C. R., Ríos Touma, B. P., Rieradevall i Sant, M., & Prat i Fornells, N. (2009). Propuesta de un protocolo de evaluación de la calidad ecológica de ríos andinos (CERA) y su aplicación a dos cuencas de Ecuador y Perú. Limnetica, 2009, vol. 28, núm. 1, p. 35-64.

Alba-Tercedor, J. (1996). Macroinvertebrados acuáticos y calidad de las aguas de los ríos. In IV Simposio del Agua en Andalucía (SIAGA), Almería, España (pp. 203-213).

Ambelu, A., Lock, K., & Goethals, P. (2010). Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. Ecological Informatics, 5(2), 147-152.

Andah, K. R. Rosso. A. C, Taramasso. (1987). The role of quantitative geomorphology in the hydrological response of river networks. Water for the Future: Hydrology in Perspective (Rome Symposium, April 1987). IAHS Publ. no. 164.

Art Borkent & Gustavo R. Spinelli. (2007). Neotropical Ceratopogonidae (Diptera: Insecta). In: Adis, J., Arias, J.R., Rueda-Delgado, G. & K.M. Wantzen (Eds.): Aquatic Biodiversity in Latin America (ABLA). Vol. 4. Pensoft, Sofia-Moscow, 198 pp.

Armitage, P. D., Moss, D., Wright, J. F., & Furse, M. T. (1983). The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water research, 17(3), 333-347.

Arteaga, F., & Ferrer, A. (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. Journal of chemometrics, 16(8-10), 408-418.

Astudillo S., P. Astudillo Webster, P. Cisneros, C. Coello, J. García, C. González, P. Lazo, E. Pacheco, A. Rengel, B. Stoop, S. Van Noten, A. Wijffels, A. Zúñiga. (2010). Atlas de la Cuenca del Río Paute. PROMAS - Universidad de Cuenca.

Bailey, T., & Jain, A. K. (1978). A Note on Distance-Weighted $ k $-Nearest Neighbor Rules. IEEE Transactions on Systems, Man, and Cybernetics, (4), 311-313.

Bain, M. B., & Stevenson, N. J. (1999). Aquatic habitat assessment. Asian Fisheries Society, Bethesda.

Barbour, M. T., Gerritsen, J., Snyder, B. D., & Stribling, J. B. (1999). Rapid bioassessment protocols for use in streams and wadeable rivers. USEPA, Washington.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods, 5(16), 3790-3798.

Bere, T., & Tundisi, J. G. (2011). The effects of substrate type on diatom-based multivariate water quality assessment in a tropical river (Monjolinho), São Carlos, SP, Brazil. Water, Air, & Soil Pollution, 216(1-4), 391-409.

Beschta, R. L., Bilby, R. E., Brown, G. W., Holtby, L. B., & Hofstra, T. D. (1987). CHAPTER SIX Stream Temperature and Aquatic Habitat: Fisheries and Forestry Interactions.

Beschta, R. L. (1997). Riparian shade and stream temperature: an alternative perspective. Rangelands, 25-28.

Bispo, P. C., Oliveira, L. G., Bini, L. M., & Sousa, K. G. (2006). Ephemeroptera, Plecoptera and Trichoptera assemblages from riffles in mountain streams of Central Brazil: environmental factors influencing the distribution and abundance of immatures. Brazilian Journal of Biology, 66(2B), 611-622.

Bonada, N., Rieradevall, M., Prat, N., & Resh, V. H. (2006). Benthic macroinvertebrate assemblages and macrohabitat connectivity in Mediterranean-climate streams of northern California. Journal of the North American Benthological Society, 25(1), 32-43.

Bonada, N., Rieradevall, M., & Prat, N. (2007). Macroinvertebrate community structure and biological traits related to flow permanence in a Mediterranean river network. Hydrobiologia, 589(1), 91-106.

Bojsen, B. H., & Barriga, R. (2002). Effects of deforestation on fish community structure in Ecuadorian Amazon streams. Freshwater Biology, 47(11), 2246-2260.

Boulton, A. J. (2003). Parallels and contrasts in the effects of drought on stream macroinvertebrate assemblages. Freshwater Biology, 48(7), 1173-1185.

Cadima, J., & Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. Journal of Applied Statistics, 22(2), 203-214.

Carrera, C. & Fierro, K. (2001). Manual de monitoreo: los macroinvertebrados acuáticos como indicadores de la calidad del agua. EcoCiencia. Quito.

Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith, V. H. (1998). Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecological applications, 8(3), 559-568.

Chapman, D. (1992) Water Quality Assessment, ed. D. Chapman on behalf of UNESCO, WHO and UNEP, 585pp. Chapman & Hall, London.

Chaves, M. L., Rieradevall, M., Chainho, P., Costa, J. L., Costa, M. J., & Prat, N. (2008). Macroinvertebrate communities of non-glacial high altitude intermittent streams. Freshwater Biology, 53(1), 55-76.

Chessman, B. C. (1995). Rapid assessment of rivers using macroinvertebrates: A procedure based on habitat-specific sampling, family level identification and a biotic index. Australian Journal of Ecology, 20(1), 122-129.

Chutter, F. M. (1972). An empirical biotic index of the quality of water in South African streams and rivers. Water Research, 6(1), 19-30.

Coscarón Sixto & Cecilia L. Coscarón Arias. (2007). Neotropical Simuliidae (Diptera: Insecta). In: Adis, J., Arias, J.R., Rueda-Delgado, G. & K.M. Wantzen (Eds.): Aquatic Biodiversity in Latin America (ABLA). Vol. 3. Pensoft, Sofia-Moscow, 685 pp.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1), 21-27.

Crawford, J. K., & Lenat, D. R. (1989). Effects of land use on the water quality and biota of three streams in the Piedmont Province of North Carolina.

Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. Mult Classif Syst, 1-17.

Dauwalter, D. C., Splinter, D. K., Fisher, W. L., & Marston, R. A. (2007). Geomorphology and stream habitat relationships with smallmouth bass (Micropterus dolomieu) abundance at multiple spatial scales in eastern Oklahoma. Canadian Journal of Fisheries and Aquatic Sciences, 64(8), 1116-1129.

De Ceballos, B. S. O., König, A., & De Oliveira, J. F. (1998). Dam reservoir eutrophication: a simplified technique for a fast diagnosis of environmental degradation. Water Research, 32(11), 3477-3483.

DECAMPS, H. (1996). THE RENEWAL OF FLOODPAIN FORESTS ALONG RIVERS: A LANDSCAPE PERSPECTIVE. Verhandlungen-Internationale Vereinigung für theoretische und angewandte Limnologie, 26, 35-59.

D'heygere, T., Goethals, P. L., & De Pauw, N. (2003). Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecological Modelling, 160(3), 291-300.

Díaz, F. R., López, F. J. B., Font, E. S., & Guijosa, L. P. (2010). Bioestadística: métodos y aplicaciones.

Dixon, W., & Chiswell, B. (1996). Review of aquatic monitoring program design. Water research, 30(9), 1935-1948.

Dollar, E. S. J., James, C. S., Rogers, K. H., & Thoms, M. C. (2007). A framework for interdisciplinary understanding of rivers as ecosystems. Geomorphology, 89(1), 147-162.

Domínguez-Granda, L., Lock, K., & Goethals, P. L. (2011). Using multi-target clustering trees as a tool to predict biological water quality indices based on benthic macroinvertebrates and environmental parameters in the Chaguana watershed (Ecuador). Ecological Informatics, 6(5), 303-308.

Domínguez-Granda, L., Goethals, P., & De Pauw, N. (2005). Aspectos del ambiente físico-químico del río Chaguana: un primer paso en el uso de los macroinvertebrados bentónicos en la evaluación de su calidad de agua. Revista Tecnológica-ESPOL, 18(1).

Domínguez, E. & H. R. Fernández (Eds.). (2009), Macroinvertebrados bentónicos sudamericanos. Sistemática y biología. Fundación Miguel Lillo, Tucumán, Argentina. 656 pp.

Domínguez, E., Molineri, C., Pescador, M.L., Hubbard, M.D. & C. Nieto. (2006). Ephemeroptera of South America. In: Adis, J., Arias, J.R., Rueda-Delgado, G. & K.M. Wantzen (Eds.): Aquatic Biodiversity in Latin America (ABLA). Vol. 2. Pensoft, Sofia-Moscow, 646 pp.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. Systems, Man and Cybernetics, IEEE Transactions on, (4), 325-327.

Dudgeon, D. (1996). Life histories, secondary production, and microdistribution of heptageniid mayflies (Ephemeroptera) in a tropical forest stream. Journal of Zoology, 240(2), 341-361.

Dudgeon, D. (1998). The population dynamics of three species of Calamoceratidae (Trichoptera) in a tropical forest stream. In Abstract In: 9th International Symposium on Trichoptera. Chiang Mai University, Thailand.

Dudgeon, D. (Ed.). (2011). Tropical stream ecology. Academic Press.

Eaton, L. E., & Lenat, D. R. (1991). Comparison of a rapid bioassessment method with North Carolina's qualitative macroinvertebrate collection method. Journal of the North American Benthological Society, 335-338.

Fattorelli, S., & Fernández, P. C. (2007). Diseño hidrológico. Zeta Editores.

Feminella, J. W. (1996). Comparison of benthic macroinvertebrate assemblages in small streams along a gradient of flow permanence. Journal of the North American Benthological Society, 651-669.

Fisher, S. G., & Likens, G. E. (1973). Energy flow in Bear Brook, New Hampshire: an integrative approach to stream ecosystem metabolism. Ecological monographs, 43(4), 421-439.

Flecker, A. S., & Feifarek, B. (1994). Disturbance and the temporal variability of invertebrate assemblages in two Andean streams. Freshwater Biology, 31(2), 131-142.

Buss, D. F., & Salles, F. F. (2007). Using Baetidae species as biological indicators of environmental degradation in a Brazilian river basin. Environmental Monitoring and Assessment, 130(1-3), 365-372.

Frank, I. E., & Todeschini, R. (1994). The data analysis handbook. Elsevier.

Frothingham, K. M., Rhoads, B. L., & Herricks, E. E. (2002). A multiscale conceptual framework for integrated ecogeomorphological research to support stream naturalization in the agricultural Midwest. Environmental Management, 29(1), 16-33.

Gallegos Sánchez, S. A. (2013). Effect of riparian vegetation cover and season on aquatic macroinvertebrate assemblages in the Ecuadorian Andes.

Giller, P. S., & Malmqvist, B. (1998). The biology of streams and rivers.

Girel, J., & Manneville, O. (1998). Present species richness of plant communities in alpine stream corridors in relation to historical river management. Biological Conservation, 85(1), 21-33.

Gonze, D. (2007). Principal Components Analysis.

Juahir, H., Zain, S. M., Aris, A. Z., Yusoff, M. K., & Mokhtar, M. B. (2010). Spatial assessment of Langat river water quality using chemometrics. Journal of Environmental Monitoring, 12(1), 287-295.

Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4(1): 9pp. http://palaeo-electronica.org/2001_1/past/issue1_01.htm

Hand, D. J. (2012). Assessing the performance of classification methods. International Statistical Review, 80(3), 400-414.

Hannah, D. M., Sadler, J. P., & Wood, P. J. (2007). Hydroecology and ecohydrology: a potential route forward?. Hydrological Processes, 21(24), 3385-3390.

Hawkins, C. P., Kershner, J. L., Bisson, P. A., Bryant, M. D., Decker, L. M., Gregory, S. V., ... & Young, M. K. (1993). A hierarchical approach to classifying stream habitat features. Fisheries, 18(6), 3-12.

He, H., Graco, W., & Yao, X. (1999). Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. In Simulated Evolution and Learning (pp. 74-81). Springer Berlin Heidelberg.

Heckman, C. W. (2008). Encyclopedia of South American aquatic insects: Odonata-Zygoptera: Illustrated keys to known families, genera, and species in South America. Springer Science & Business Media.

Heckman, C. W. (2011). Encyclopedia of South American Aquatic Insects: Hemiptera-Heteroptera: Illustrated Keys to Known Families, Genera, and Species in South America. Springer Science & Business Media.

Hilsenhoff, W. L. (1982). Using a biotic index to evaluate water quality in streams (pp. 1-22). Madison, Wisconsin: Department of Natural Resources.

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. Water resources research, 18(1), 107-121.

Hoang, T. H., Lock, K., Mouton, A., & Goethals, P. L. (2010). Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Ecological Informatics, 5(2), 140-146.

Ison, M., Sitt, J., & Trevisan, M. (2005). Algoritmos genéticos: aplicación en MATLAB.

Jacobsen, D. (2004). Contrasting patterns in local and zonal family richness of stream invertebrates along an Andean altitudinal gradient. Freshwater Biology, 49(10), 1293-1305.

Jáimez-Cuéllar, P., Vivas, S., Bonada, N., Robles, S., Mellado, A., Álvarez, M., ... & Alba-Tercedor, J. (2002). Protocolo GUADALMED (prece). Limnetica, 21(3-4), 187-204.

Jiang, Y., & Zhou, Z. H. (2004). Editing training data for knn classifiers with neural network ensemble. In Advances in Neural Networks–ISNN 2004 (pp. 356-361). Springer Berlin Heidelberg.

Jóźwik, A. (1983). A learning scheme for a fuzzy k-NN rule. Pattern Recognition Letters, 1(5), 287-289.

Quinn, J. M., & Hickey, C. W. (1990). Characterisation and classification of benthic invertebrate communities in 88 New Zealand rivers in relation to environmental factors. New Zealand journal of marine and freshwater research, 24(3), 387-409.

Kannel, P. R., Lee, S., Kanel, S. R., & Khan, S. P. (2007). Chemometric application in classification and assessment of monitoring locations of an urban river system. Analytica Chimica Acta, 582(2), 390-399.

Karr, J. R., Fausch, K. D., Angermeier, P. L., Yant, P. R., & Schlosser, I. J. (1986). Assessing biological integrity in running waters. A method and its rationale. Illinois Natural History Survey, Champaign, Special Publication, 5.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. Systems, Man and Cybernetics, IEEE Transactions on, (4), 580-585.

Kendall, M. (1980). Multivariate analysis. Second edition. Charles Griffin, London, UK.

Knox M. V. (1968). Two new species of the Genus *Baetodes* from Ecuador (Ephemeroptera: Baetidae). The Pan-Pacific Entomologist, 44(3), 251-257.

Koklu, R., Sengorur, B., & Topal, B. (2010). Water quality assessment using multivariate statistical methods—a case study: Melen River System (Turkey). Water resources management, 24(5), 959-978.

Kowalkowski, T., Zbytniewski, R., Szpejna, J., & Buszewski, B. (2006). Application of chemometrics in river water classification. Water Research, 40(4), 744-752.

Kuczera, G., & Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. Water Resources Research, 34(6), 1481-1489.

Kutzer, C. (2008). Potential of the Knn method for estimation and monitoring off-reserve forest resources in Ghana (Doctoral dissertation, Faculty of Forest and Environmental Sciences, Albert-Ludwigs-Universität Freiburg).

Leopold, L. B., Wolman, M. G., & Miller, J. P. (2012). Fluvial processes in geomorphology. Courier Corporation.

Lewis Jr, W. M., Hamilton, S. K., & Saunders III, J. F. (1995). Rivers of northern South America. Ecosystems of the world: Rivers, 219-256.

Librando, V. (1991). Chemometric evaluation of surface water quality at regional level. Fresenius' Journal of Analytical Chemistry, 339(9), 613-619.

Loinaz, M. C., Bauer-Gottwein, P., & Butts, M. (2012). Integrated ecohydrological modeling at the catchment scale. DHI DenmarkDHI Denmark.

López-Luna, J., Ibáñez, M. A., & Villarroel, M. (2013). Using multivariate analysis of water quality in RAS with Nile tilapia (Oreochromis niloticus) to model the evolution of macronutrients. Aquacultural Engineering, 54, 22-28.

Lowrance, R., Altier, L. S., Newbold, J. D., Schnabel, R. R., Groffman, P. M., Denver, J. M., ... & Todd, A. H. (1997). Water quality functions of riparian forest buffers in Chesapeake Bay watersheds. Environmental Management, 21(5), 687-712.

Mandaville, S. M. (2002). Benthic macroinvertebrates in freshwaters: Taxa tolerance values, metrics, and protocols (Vol. 128, p. 315). Halifax, Canada: Soil & Water Conservation Society of Metro Halifax.

Maddock, I., Harby, A., Kemp, P. and Wood, P. (2013) Ecohydraulics: An Introduction, in Ecohydraulics: An Integrated Approach (eds I. Maddock, A. Harby, P. Kemp and P. Wood), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118526576.ch1

McCune, B., Grace, J. B., & Urban, D. L. (2002). Analysis of ecological communities (Vol. 28). Gleneden Beach, Oregon: MjM software design.

Molina, E. (2008). El Consejo de Gestión de Aguas de la cuenca del Paute. Experiencias y lecciones. Seminario Internacional" Cogestión de Cuencas Hidrográficas Experiencias y Desafíos". Turrialba (Costa Rica). 14-16 Oct 2008.

Mustapha, A., & Abdu, A. (2012). Application of Principal Component Analysis & Multiple Regression Models in Surface Water Quality Assessment. Journal of Environment and Earth Science, 2(2), 16-23.

Niemann, H. (1983). Klassifikation von mustern (Vol. 1). Berlin: Springer.

Osborne, L. L., & Kovacic, D. A. (1993). Riparian vegetated buffer strips in water-quality restoration and stream management. Freshwater biology, 29(2), 243-258.

Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. Water research, 39(12), 2621-2635.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2003). Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology, 84(9), 2347-2363.

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, 81(3), 719-745.

Plafkin, J. L., Barbour, M. T., Porter, K. D., Gross, S. K., & Hughes, R. M. (1989). Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. In Rapid bioassessment protocols for use in streams and rivers: Benthic macroinvertebrates and fish. EPA.

Pringle, C. M., Paaby-Hansen, P., Vaux, P. D., & Goldman, C. R. (1986). In situ nutrient assays of periphyton growth in a lowland Costa Rican stream. Hydrobiologia, 134(3), 207-213.

Ramírez, A., & Pringle, C. M. (1998). Structure and production of a benthic insect assemblage in a neotropical stream. Journal of the North American Benthological Society, 443-463.

Rankin, E. T. (1995). Habitat indices in water resource quality assessments. Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. CRC Press, Boca Raton, FL, 181-208.

Raven, P. J. (1998). River Habitat Quality: the physical character of rivers and streams in the UK and Isle of Man. Environment Agency.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). Applied regression analysis: a research tool. Springer Science & Business Media.

Reisenhofer, E., Adami, G., & Barbieri, P. (1998). Using chemical and physical parameters to define the quality of karstic freshwaters (Timavo River, North-Eastern Italy): A chemometric approach. Water Research, 32(4), 1193-1203.

Rincón, J., & Cressa, C. (2000). Temporal variability of macroinvertebrate assemblages in a neotropical intermittent stream in Northwestern Venezuela. Archiv für Hydrobiologie, 148(3), 421-432.

Ríos-Touma, B., Acosta, R., & Prat, N. (2006). Distribution of macroinvertebrate communities in the high Andes and their tolerance to pollution. A review and proposal of a biotic index for high Andean streams (Andean Biotic Index, ABI).

Ríos-Touma, B., Acosta, R., & Prat, N. (2014). The Andean Biotic Index (ABI): revised tolerance to pollution values for macroinvertebrate families and index performance evaluation. Revista de Biología Tropical, 62, 249-273.

Rodríguez, F. F., Duarte, R. M., & Marquínez, J. (1997). Aplicación de un Sistema de Información Geográfica en la cartografía temática y clasificación geomorfológica de los sistemas fluviales en Asturias. La Revista de la Sociedad Geológica de España, 10(1-2), 117-130.

Roldan, G. (1996). Guía para el estudio de los macroinvertebrados acuáticos del Departamento de Antioquia, Primera Edición, Editorial Impreades Presencia S.A., Bogotá, Colombia. 217 pp.

Roldán, G. (2003). Bioindicación de la calidad del agua en Colombia. Propuesta para el uso del método BMWP/Col. Editorial Universidad de Antioquia. Colección de Ciencia y Tecnología. Medellín.

Rosgen, D. L. (1985, April). A stream classification system. In Riparian ecosystems and their management: reconciling conflicting uses. First North American Riparian Conference, Arizona (pp. 91-95).

Rosgen, D. L. (1994). A classification of natural rivers. Catena, 22(3), 169-199.

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J. Cheminformatics, 5, 27.

Schade, J. D., Marti, E., Welter, J. R., Fisher, S. G., & Grimm, N. B. (2002). Sources of nitrogen to the riparian zone of a desert stream: implications for riparian vegetation and nitrogen retention. Ecosystems, 5(1), 68-79.

Schade, J., Fisher, S. G., Grimm, N. B., & Seddon, J. A. (2001). The influence of a riparian shrub on nitrogen cycling in a Sonoran Desert stream. Ecology, 82(12), 3363-3376.

Sharaf, M. A. (1986). Chemometrics (Vol. 82). John Wiley & Sons.

Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environmental Modelling & Software, 22(4), 464-475.

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., ... & Kouimtzis, T. (2003). Assessment of the surface water quality in Northern Greece. Water research, 37(17), 4119-4124.

Sotomayor. G. (2007). Análisis de los efectos causados por la actividad minera sobre los cuerpos de agua en la microcuenca del Río Tenguel, cantones Ponce Enríquez y Pucará, mediante la utilización de parámetros físico - químicos y biológicos. Universidad del Azuay – Escuela de Biología.

Stark, B. P. (2007). *Anacroneuria marshalli* (Plecoptera: Perlidae), a new stonefly from Argentina, and two new records from Ecuador. Illiesia, 3(17), 171-173.

Strobl, R. O., & Robillard, P. D. (2008). Network design for water quality monitoring of surface freshwaters: A review. Journal of environmental management, 87(4), 639-648.

Sullivan K. & Adams T. (1991). The physics of stream heating: an analysis of temperature patterns in stream environments based on physical principles and field data. Weyerheauser Technical Report. 044–5002 • 90/2. Technology Center, Tacoma, WA 98477.

Suguna, N., & Thanushkodi, K. (2010). An improved K-nearest neighbor classification using Genetic Algorithm. International Journal of Computer Science Issues, 7(2), 18-21.

Suárez, M. L., Vidal-Abarca Gutiérrez, M. R., Sánchez-Montoya, M. D. M., Alba-Tercedor, J., Álvarez, M., Avilés, J., ... & Vivas, S. (2002). Las riberas de los ríos mediterráneos y su calidad: el uso del índice QBR. Limnetica, 2002, vol. 21, num. 3-4, p. 135-148.

Tarboton, D. G., Bras, R. L., & Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. Hydrological processes, 5(1), 81-100.

Tauler, R., Walczak, B., & Brown, S. D. (2009). Comprehensive chemometrics: chemical and biochemical data analysis. Elsevier.

Townsend, S. A., & Padovan, A. V. (2005). The seasonal accrual and loss of benthic algae (Spirogyra) in the Daly River, an oligotrophic river in tropical Australia. Marine and Freshwater Research, 56(3), 317-327.

Todeschini, R. (1998). Introduzione alla chemiometria. EdiSES, Napoli, 321.

Thoms, M. C., & Parsons, M. E. L. I. S. S. A. (2002). Eco-geomorphology: an interdisciplinary approach to river science. International Association of Hydrological Sciences, Publication, (276), 113-119.

Varmuza, K. (1980). Pattern recognition in analytical chemistry. Analytica Chimica Acta, 122(3), 227-240.

Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water research, 32(12), 3581-3592.

Voelz, N. J., & McArthur, J. V. (2000). An exploration of factors influencing lotic insect species richness. Biodiversity & Conservation, 9(11), 1543-1570.

Vogel, S. (1996). Life in moving fluids: the physical biology of flow. Princeton University Press.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

Ward, J. V. (1998). Riverine landscapes: biodiversity patterns, disturbance regimes, and aquatic conservation. Biological conservation, 83(3), 269-278.

Webb, B. W., & Zhang, Y. (1997). Spatial and seasonal variability in the components of the river heat budget. Hydrological Processes, 11(1), 79-101.

Williams, D. D., & Feltmate, B. W. (1992). Aquatic insects. CAB international.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. Systems, Man and Cybernetics, IEEE Transactions on, (3), 408-421.

Woodiwiss, F. (1964). The biological system of stream classification used by the Trent River Board. Chemistry and Industry 14, 443-447.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1), 37-52.

Wunderlin, D.A., Diaz, M.P., Ame, M.V., Pesce, S.F., Hued, A.C., Bistoni, M.A., (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). Water research, 35(12), 2881-2894.

Zalewski, M. (2002). Ecohydrology—the use of ecological and hydrological processes for sustainable management of water resources/Ecohydrologie—la prise en compte de processus écologiques et hydrologiques pour la gestion durable des ressources en eau. Hydrological Sciences Journal, 47(5), 823-832.

Zalewski, M., Janauer, G. A., & Jolankai, G. (1997). Ecohydrology: a new paradigm for the sustainable use of aquatic resources. Ecohydrology. A New Paradigm for the Sustainable Use of Aquatic Resources.

Zwanziger, J. W. E. H. W., & Geiß, S. (1997). Chemometrics in environmental analysis. Wiley-VCH.